

Prediction of Hydrate and Solvate
Formation Using Knowledge-based Models

Khaled Takieddin



Thesis submitted for the degree of Doctor of Philosophy

School of Pharmacy, University of East Anglia

September 2016

Khaled Takieddin

2016

Prediction of Hydrate and Solvate Formation Using Knowledge-based Models

Abstract

Solvate formation is a phenomenon that has received special attention in solid state chemistry over the past few years. This is due to its potential to both improve and impair pharmaceutical formulations. The reasons for solvate formation aren't explicitly known. Therefore, there is currently no reliable guide in the literature on what solvents to choose in order to avoid or form a solvate when crystallizing an organic material. In this thesis we address the problem by trying to find the main reasons of solvate formation. A knowledge-based approach was used to link the molecular structure of an organic compound to its ability to form a solvate with five different solvents; these are ethanol, methanol, dichloromethane, chloroform and water. The Cambridge Structural Database (CSD) was used as a source of information for this study. A supervised machine learning method, logistic regression was found to be the optimal method for fitting these knowledge-based models. The result was one predictive model per solvent, with a success rate of 74-80 %. Each model incorporated two molecular descriptors, representing two molecular features of molecules. These are the size and branching in addition to hydrogen bonding ability. The models' predictive ability was validated *via* experimental work, in which slurries of 10 pharmaceutically active ingredients were screened for solvate formation with each of the five solvents in the study. During the screening process, a new diflunisal dichloromethane solvate, a diflunisal chloroform solvate and a hymercromone methanol solvate were found. The PXRD patterns of these forms are reported. The thesis also includes SCXRD analysis of a previously known grisoefulvin dichloromethane solvate, a previously known fenofibrate polymorph and a new fenofibrate polymorph.

To my parents

Contents:

Chapter 1:	Introduction	1
1.1	References	4
Chapter 2:	Literature review	5
2.1	Solid forms of pharmaceuticals	6
2.1.1	Amorphous materials.....	7
2.1.2	Polymorphism	8
2.1.3	Multicomponent crystalline solids.....	10
2.2	Identification of the solid form	13
2.2.1	Experimental identification of solid form	13
2.2.2	Theoretical prediction of possible solid forms.....	14
2.2.3	Previous CSD investigations on hydrate and solvate formation.....	18
2.3	Statistics	21
2.3.1	Hypothesis testing.....	22
2.3.2	Data mining and machine learning	23
2.3.3	Model selection.....	32
2.3.4	Graphical illustrations	35
2.4	Non-covalent interactions	38
2.4.1	Hydrogen bond	38
2.4.2	Halogen bond	41
2.4.3	Interactions of aromatic rings.....	41
2.5	Characterization.....	43
2.5.1	X-ray diffraction techniques.....	43
2.5.2	Thermal analysis.....	47
2.6	References	49
Chapter 3:	Materials and methods	68
3.1	Materials	69
3.2	Methods	70
3.2.1	Descriptor calculation	70
3.2.2	Slurry preparation	79
3.2.3	Thermogravimetric analysis	82
3.2.4	Single crystal X-ray diffraction.	82
3.2.5	PXRD	83
3.2.6	Microscopy.....	83
3.3	References	84
Chapter 4:	Data acquisition and descriptor calculation	89

4.1	Overview of the research steps	90
4.2	Data collection	90
4.2.1	Solvent selection rationale.....	90
4.2.2	Entries selection rationale	92
4.2.3	Refinement of the non-solvate/solvate forming groups	94
4.2.4	Further partitioning of the solvate group.....	98
4.2.5	Preparing the data for descriptor calculation.....	100
4.3	Descriptor calculation	102
4.3.1	Entries with calculation problems.....	103
4.3.2	Descriptors with calculation problems	105
4.4	Summary of data.....	105
4.5	References	112
Chapter 5:	Statistical analysis	114
5.1	Overview	115
5.2	Dimensionality reduction.....	116
5.2.1	The Wilcoxon rank sum test.....	116
5.2.2	Principal component analysis	120
5.3	Supervised machine learning.....	128
5.3.1	Selection of the machine learning algorithm.....	128
5.3.2	Choosing descriptors to decide the linearity of the problem	129
5.3.3	Equal size sampling	130
5.3.4	Parameter adjustment of the RBF kernel	132
5.3.5	SVM vs LR	133
5.4	Principal components as logistic regression variables.....	136
5.4.1	Models with one principal component.....	137
5.4.2	Models with two and three principal components	139
5.5	Systematic variable selection using logistic regression	142
5.5.1	Single-variable models	142
5.5.2	Two variable models	144
5.5.3	Three-variable models	148
5.6	A closer look on the two-variable models	152
5.6.1	The models.....	153
5.6.2	The meaning of the descriptors	156
5.6.3	The descriptor values and their coefficients.....	159
5.6.4	Visual representation.....	161
5.6.5	Cut-off point determination.....	164
5.6.6	Residuals	166

5.6.7	Misclassified data and intercept adjustment.....	173
5.7	Simple alternatives.....	175
5.7.1	The alternative descriptors.....	176
5.7.2	Performance of the simple models.....	179
5.8	Practical usage of the models.....	181
5.8.1	Mathematical representation.....	181
5.8.2	Visual representation.....	185
5.9	References.....	187
Chapter 6:	Discussion of the models.....	191
6.1	Effects the models take into account.....	192
6.1.1	Size and branching.....	192
6.1.2	Hydrogen bonding.....	197
6.2	Effects the models do not take into account.....	203
6.2.1	Hydrogen bond strengths.....	204
6.2.2	Accessibility of hydrogen bonding.....	206
6.2.3	Ring interactions.....	212
6.2.4	Halogen bonding.....	215
6.2.5	Zwitterions.....	220
6.2.6	Why these factors were not included in the predictive models.....	225
6.3	Possible improvements.....	226
6.3.1	Inclusion of a hydrogen bond strength scale.....	226
6.3.2	Application of Etter's rules.....	228
6.3.3	Inclusion of steric hindrance description.....	229
6.4	References.....	230
Chapter 7:	Experimental validation of the models.....	241
7.1	Overview.....	242
7.2	Selection of drug candidates and their profiles.....	242
7.3	Sample preparation and characterization.....	253
7.4	Prediction vs results.....	270
7.5	Griseofulvin dichloromethane solvate.....	276
7.5.1	Overview.....	276
7.5.2	Under the microscope.....	276
7.5.3	X-ray data and structure solution.....	277
7.5.4	Interactions and packing.....	283
7.6	References.....	287
Chapter 8:	Single crystal analysis of the new fenofibrate forms.....	294
8.1	Overview.....	295

8.2	Fenofibrate polymorphs	295
8.3	Polymorph preparation.....	296
8.4	Single crystal X-ray Diffraction	298
8.4.1	Structure solution	298
8.4.2	Main intermolecular interactions	302
8.4.3	Comparison of crystal structures of fenofibrate form I, form IIa and form III..	309
8.5	References	313
Chapter 9:	Summary and conclusions	314
9.1	Summary	315
9.2	Conclusions and future outlook.....	320
9.2.1	Conclusions	320
9.2.2	Future outlook	322

List of Figures:

Figure 2-1. Graphical illustration of Gibbs energy in monotriopic polymorph system (A) and in enantiotropic polymorph system (B). M.p.is melting point of Form 1 or Form 2 correspondingly and T.p. is polymorph transition point in enantiotropic system. ²²	9
Figure 2-2. (a) A distribution with a left-tailed alpha value and an equivalent p-value . (b) A distribution with a right-tailed alpha value and an equivalent p-value . (c) A distribution with a two-tailed alpha value and a p-value equivalent to alpha/2.	23
Figure 2-3. An illustration of 4 variables that point in different directions (have some correlation) in terms of principal components 1 and 2	25
Figure 2-4 Values of Y (on the y-axis) vs the values of the continuous variable X (on the x axis)	27
Figure 2-5. Fitted probabilities (Y axis) vs the values of X (X axis). Red and black points correspond to two classes, representing a case of binary data.	28
Figure 2-6. An illustration of support vector machine linearly separating binary data.....	29
Figure 2-7. An illustration of non-linear vector machine with separating binary data.	30
Figure 2-8. An illustration of non-linear support vector machine with soft margins, note that the algorithm converged despite the misclassification of one black point.	31
Figure 2-9. An illustration of how a 5-fold cross-validation works.	33
Figure 2-10. An example of the ROC curve of the chloroform model presented in this thesis..	36
Figure 2-11. Sensitivity and specificity curves, with the optimal threshold level shown in the dotted line.....	37
Figure 2-12. An example of a boxplot, obtained as a comparison between the solvate and the non-solvate groups in this thesis.	38
Figure 2-13. Graphic representation of ring interaction types.....	42
Figure 2-14. A comparison of desolvation events in TG thermograms for stoichiometric and non-stoichiometric solvates.....	48
Figure 3-1. Molecular graph of the aspirin molecule.....	74
Figure 4-1. Main steps of analysis of solvate formation.....	90
Figure 4-2. The number of solvates recorded in CSD for the 10 most commonly used organic recrystallization solvents.	91
Figure 4-3. Percentages of entries with unknown recrystallization solvent among the solvate-forming entries in each solvent dataset.	95
Figure 4-4. The CSD entry CAWREY showing more than one solvent in the crystal structure. ...	97
Figure 4-5. An illustration of the four individual groups of each solvent.	98

Figure 4-6. The number of entries that were recrystallized from the solvent of interest and the number of entries recrystallized from a mixture as obtained from the database. Note: numbers were rounded to the nearest integer this is why they do not all sum precisely to 100 %.	99
Figure 4-7. The splitting step of the AFEYOA12 (a) and the COHLOC13 (b) entries.	101
Figure 4-8. QUPKIV molecule and how it is recorded in the CSD (left) and the auto-edited structure by Mercury (right).	103
Figure 4-9. A zwitterionic molecule from the dichloromethane non-solvate dataset. The WIZMAU molecule.	105
Figure 4-10. Percentage of solvate and non-solvate forming molecules in each solvent's dataset. Total number of entries is 19,010.	106
Figure 4-11. The chemical space covered by each solvent dataset (units in Da (g/mol)). Y axis shows the frequency of occurrences.	107
Figure 4-12. Log P values for each solvent dataset. Y axis shows the frequency of occurrences.	108
Figure 4-13. The number of hydrogen bond donors in each solvent's dataset. Y axis shows the frequency of occurrences.	109
Figure 4-14. The number of hydrogen bond acceptors in each solvent's dataset. Y axis shows the frequency of occurrences.	110
Figure 5-1. The thinking process throughout Chapter 5.	115
Figure 5-2. Boxplot of the nAT descriptor value in the ethanol dataset (Y axis). The circles represent the outliers in the dataset.	118
Figure 5-3. Insignificant difference between the groups in the O % descriptor. Such descriptors were omitted from the dataset.	119
Figure 5-4. An illustration of the number of the variables that were omitted due to insignificance between the solvate-forming and the non-solvate forming groups. Total number of descriptors in each circle is 4885.	120
Figure 5-5. The percentage of Variance explained by the top 100 principal components in the ethanol dataset. Each bar represents a principal component, ordered.	122
Figure 5-6. The ethanol data points in terms of the first three principal components. The solvate-forming molecules are shown in blue and the non-solvate forming molecules are shown in red.	123
Figure 5-7. Pairs plot of ethanol data points in terms of PC1, PC2 and PC3. The solvate-forming molecules are shown in blue and the non-solvate forming molecules are shown in red.	124
Figure 5-8. Histograms of the rotation values of the first (a), second (b) and third (c) principal components in the ethanol's dataset. Total number of variables is 2647. The colours of the bars have no indication, they are used for a better illustration.	127

Figure 5-9. An example of a KNIME workflow illustrating the steps of analysis that were taken to compare SVM to LR. The CSV Reader node reads the data. The Column Filter node selects what variables will be included in the model fitting. The Equal Size Sampling node takes a sample of the data in which the two classes (solvate and nonsolvate) are equal in number. The Partitioning data node splits the data into a training set (10 % of the data) and a test set (90 % of the data). The Learner and Predictor nodes fits the model and performs the prediction, respectively. The Scorer node gives the % accuracy (percentage of correct predictions).	131
Figure 5-10. KNIME workflow for optimizing SVM parameters, testing an SVM model and testing a logistic regression model. This was done for 3 samples per variable which gives 9 trials for each of the solvents, resulting 45 similar workflows.	133
Figure 5-11. The percentage of correct predictions from the ethanol dataset made by the logistic regression and the support vector machine models that were fitted to the same samples. The test set is more than 1200 molecules.	134
Figure 5-12. The percentage of correct predictions from the methanol dataset made by the logistic regression and the support vector machine models that were fitted to the same samples. The test set is more than 2700 molecules.	134
Figure 5-13. The percentage of correct predictions from the dichloromethane dataset made by the logistic regression and the support vector machine models that were fitted to the same samples. Test set is more than 2300 molecules.	135
Figure 5-14. The percentage of correct predictions from the chloroform dataset made by the logistic regression and the support vector machine models that were fitted to the same samples. Test set is more than 2100 molecules.	135
Figure 5-15. The percentage of correct predictions from the water dataset made by the logistic regression and the support vector machine models that were fitted to the same samples. Test set is more than 500 molecules.	136
Figure 5-16. Receiver operative curve (ROC) of the ethanol model 1. The area under the curve is approx. 0.72. The dotted line, with the 45 ° angle represents the random guess (50 % correct prediction).....	139
Figure 5-17. ROC curves of the water models taking into account 1, 2 and 3 principal components. The curves almost overlap, showing the insignificance of the addition of the second and the third principal components.	141
Figure 5-18. Reduction of the MSE by the addition of the second descriptor in each solvent.	148
Figure 5-19. A plot of the normalized AIC and MSE value for the best and the worst two-variable models. The details of these models were shown in Table 5-9.	150

Figure 5-20. The change in MSE between the two-variable and the three-variable models. For each solvent, the models with the lowest and the highest MSE of the 10 equally sized samples is shown.	152
Figure 5-21. A plot of 2600 datapoints from the dichloromethane dataset in terms of the 2 descriptors that give the best linear separation, these are the SM3_H2 and Hy. The black line represents the decision boundary upon which the outcome is predicted. Colours = experimental outcome.....	161
Figure 5-22. Histograms of the SM3_H2 descriptor distribution from the chloroform data. ..	162
Figure 5-23. An illustration of the change in performance between the alternative water model fitted using piLD and nH (red) vs the water model fitted using piLD, nH and nHDon (blue). The steps in the curve (not smooth) are due to the small number of datapoints in the water sample (number of datapoints is 606).	163
Figure 5-24. The sensitivity and specificity curves from the dichloromethane data, sample 1.	165
Figure 5-25. A residual plot of the ethanol model (from subset data no.1).....	167
Figure 5-26. The binned residual plots from the 2 variable models in each solvent (a): ethanol. Note that each plot is based on an equal size sample to show sensible distribution around the zero line. The two grey lines in each figure represents the boundaries of 95 % error assuming the model is true.....	168
Figure 5-27. Continued. The binned residual plots from the 2 variable models in each solvent (b): methanol. Note that each plot is based on an equal size sample to show sensible distribution around the zero line. The two grey lines in each figure represents the boundaries of 95 % error assuming the model is true.....	169
Figure 5-28. Continued. The binned residual plots from the 2 variable models in each solvent (c): dichloromethane. Note that each plot is based on an equal size sample to show sensible distribution around the zero line. The two grey lines in each figure represents the boundaries of 95 % error assuming the model is true.....	170
Figure 5-29. Continued. The binned residual plots from the 2 variable models in each solvent (d): chloroform. Note that each plot is based on an equal size sample to show sensible distribution around the zero line. The two grey lines in each figure represents the boundaries of 95 % error assuming the model is true.....	171
Figure 5-30. Continued. The binned residual plots from the 2 variable models in each solvent (e): water. Note that each plot is based on an equal size sample to show sensible distribution around the zero line. The two grey lines in each figure represents the boundaries of 95 % error assuming the model is true.....	172

Figure 5-31. Percentage of misclassifications in the solvate and the non-solvate groups in each model when the complete dataset was predicted using the two-variable models.	173
Figure 5-32. Percentage of misclassifications in the solvate and the non-solvate groups in each model when the complete dataset was predicted using the intercept-adjusted two-variable models.....	174
Figure 5-33. The effect of adjusting the intercept on the MSE value of each model.....	175
Figure 5-34. The increase in the MSE value when alternative models are used. Note that the comparison in shown here is between intercept-adjusted two-variable models and intercept-adjusted alternative models.	181
Figure 5-35. The structure of azithromycin. Chemical formula C ₃₇ H ₆₇ NO ₁₃	182
Figure 5-36. The chemical structure of bryostatin with the groups contributing to the TRS and nHDon descriptors highlighted.	184
Figure 5-37. The amoxicillin molecule.	185
Figure 5-38. An illustration of the graphical prediction method used by the water model. The point in the plot represents the amoxicillin molecule.....	186
Figure 6-1. Molecular structures of (a) KUHGEA and (b) XAPTOY.	193
Figure 6-2. Molecular structures of ECEKON.	194
Figure 6-3. Molecular structures of (a) SOYQON and (b) UHUWEA.	198
Figure 6-4. The importance of hydrogen bonding in solvate formation shown via the EHUKAU (a) and the QERJED (b) entries.	199
Figure 6-5. Molecular structures of (a) VUQMEA and (b) LANRUP.	204
Figure 6-6. An illustration of the DEMFUX (a) and the LOMLEF (b) entries.....	205
Figure 6-7. An illustration of the limited accessibility of the hydrogen bond donors (the oxygen atoms in yellow) in the RACMEP entry. Part (a) of the Figure shows the capped sticks model of the entry; part (b) shows the space-filling model of the same entry. The intramolecular hydrogen bonds are shown in the blue lines between atoms in part (a).	207
Figure 6-8. (a) Capped-stick and (b) space filling representation of the AZOMIL molecule. The inaccessible donors (oxygen atoms) are highlighted in yellow. The intramolecular hydrogen bonds are shown in the blue lines between atoms in part (a).	208
Figure 6-9. An illustration of the deviation that was allowed for the T-shaped interaction from 90 °.	213
Figure 6-10. A heat map showing the number of entries that from a π - π interaction, where the x-axis shows the angle and the y-axis shows the centroid to centroid distance.	213
Figure 6-11. The percentage of structures with short ring interactions in the solvate and the non-solvate groups per solvent. Note that these hits are the ones that followed the restrictions set on the angles of the interaction.	214

Figure 6-12. Molecular structure of KUWWOP.....	216
Figure 6-13. An illustration of the halogen bond in the PIGDOZ entry. It shows the role of this type of bonding in the chloroform solvate formation. The bond takes place at a distance 0.1 Å shorter than the sum of the van der Waals radii of the two atoms.	217
Figure 6-14. Molecular structure of FOBJUB. The heteroatoms are labelled with their formal charge.	220
Figure 6-15. Molecular structures of IXAYOW and QIHFAO.	227
Figure 7-1. The heating profile for theophylline and its screening products.	254
Figure 7-2. The TG thermogram of the theophylline hydrate from room temperature.	255
Figure 7-3. TGA profiles of hymecromone and its screening product.....	256
Figure 7-4. TGA profile of hymcromone methanol solvate and hydrate.....	257
Figure 7-5. PXRD patterns of the solvated forms of hymecromone.....	258
Figure 7-6.TGA profiles of isoniazid and its screening products. Note that the heating was up to 155 °C only; this is due to the low melting point of this drug candidate.....	259
Figure 7-7. TGA profiles of ethenzamide and its screening products. Note that the heating was up to 180 °C only due to the low melting point of this drug candidate.	260
Figure 7-8. TGA profiles of carbamazepine and its screening products.	261
Figure 7-9. TGA profile of carbamazepine hydrate measured from RT.....	262
Figure 7-10. TGA profiles of diflunisal and its screening products.	263
Figure 7-11. PXRD patterns of diflunisal dichloromethane and chloroform solvate forms.....	264
Figure 7-12. TGA profiles of fenofibrate and its screening products.....	265
Figure 7-13.TGA profiles of felodipine and its screening products.	266
Figure 7-14. TGA profiles of felodipine and its screening products.....	267
Figure 7-15. TGA profiles of griseofulvin and its screening products.	268
Figure 7-16. TGA profiles of the dichloromethane and chloroform solvate forms of griseofulvin.	269
Figure 7-17. The first and the last frame of the heating cycles of griseofulvin dichloromethane solvate in the reflective (a) and transmittance microscope (b). The first images (left) were taken at ambient room temperature while the final images (right) were taken at 120 °C.....	277
Figure 7-18. Comparison of X-ray diffraction patterns of griseofulvin dichloromethane solvate simulated from crystal structure (blue) and reported in literature (orange).	279
Figure 7-19. Molecular structure of griseofulvin showing ellipsoids of thermal displacement parameters. Probability of shown ellipsoids is set to 50 %.	280
Figure 7-20. An illustration of the numbering of griseofulvin.	281
Figure 7-21. (a)An overlay of the griseofulvin molecule in the new DCM solvate (red) with all reported structures of grsieofulvin in the CSD. Reference code GRISFL is shown in blue,	

GRISFL02 is shown in yellow, GRISFL03 is shown in orange, GRISFL04 is shown in green. (b) An overlay of the new griseofulvin dichloromethane solvate (red) with the exiting chloroform solvate (grey), CSD reference code MATZEO	282
Figure 7-22. The interactions between griseofulvin molecules. Part (a) shows the dimers formed by griseofulvin molecules along the B axis. Part (b) shows four pairs of dimers along the a axis. Part (c) shows the horizontal interaction between the molecules in part (b), along the C axis. Part (d) shows the ring-like interaction that connects the dimers in part (b) diagonally.	284
Figure 7-23. The channels of dichloromethane in the griseofulvin dichloromethane solvate II from. Part (a) shows these channels along the a axis, part (b) show the same channels along the b axis and part (c) show the same channels along the c axis.	285
Figure 8-1. The structural formula of fenofibrate showing atom numbering scheme	295
Figure 8-2. PXRD patterns of the different fenofibrate polymorphs	296
Figure 8-3. ORTEP structure of fenofibrate forms IIa (a) and III (b).....	299
Figure 8-4. An overlay of fenofibrate form I (black), form IIa (blue), form IIb (red), form III (green).....	300
Figure 8-5. (a) The sheet that is formed parallel to the (001) plane via C–H···O interactions in Form I. (b) The offset π – π interactions between parallel chlorophenyl rings in addition to the “embrace” interaction in form I.....	303
Figure 8-6. (a) The main interactions in fenofibrate form IIa. (b) From IIa pattern that is common with form I.	305
Figure 8-7. (a) The sheet that is formed parallel to the (100) plane via C–H···O interactions in form III. (b) The keto group in C–H···O interaction with the hydrogen of benzene ring in form III.	307
Figure 8-8. Continued. (c) Illustrates the offset π – π interaction, the “embrace” interaction and the ether C–H···O interaction in form III. (d) The two molecules on the right side in part (c) from a different angle, showing the embrace type of interaction in form III, not short H–H interaction is noticed between the two molecules in part (d).	308
Figure 8-9. The unit cell and packing of fenofibrate form I (a) and (b), form IIa (c) and (d) and form III (e) and (f).....	311

List of Tables:

Table 2-1. Distances of some common hydrogen bonds.....	40
Table 2-2. The distances and binding energy in the different π - π interactions	43
Table 3-1. The 29 main blocks of descriptors calculated by Dragon. Details explaining the calculation of each descriptor can be found in the Dragon software documentation1.....	70
Table 3-2. Topological distance matrix of aspirin	75
Table 3-3. Squared topological distance matrix of aspirin	76
Table 3-4. Reciprocal squared topological distance matrix (H2) of aspirin	77
Table 3-5. Values of the descriptors mentioned in the thesis calculated for carbamazepine and acetaminophen molecules.....	78
Table 3-6. A summary of the preparation of each slurry used in the screening experiments ...	80
Table 4-1. The breakdown of the number of solvate and non-solvate entries by solvent.....	94
Table 4-2. The count of solvate and non-solvate entries after refinement.....	98
Table 4-3. A sample from the Dragon software's output. Specifically, the sample was taken from the dichloromethane non-solvate entries dataset. NA values represent an error in the calculation	102
Table 4-4. The count of solvate and non-solvate entries left after removing the errors	104
Table 5-1. The maximum rotation value in the first three components in each solvent dataset	127
Table 5-2. The confusion matrix and overall prediction accuracy of the water logistic regression model, fitted to the nC descriptor	130
Table 5-3. Details of the logistic regression models taking into account the first principle component (model 1) in different solvents. The models with the lowest MSE (best performance) and the models with the highest MSE (worst performance) among the 10 equal size samples for each solvent are shown.....	138
Table 5-4. Details of the logistic regression models taking into account the first two principal components (Model 2) in different solvents. The models with the lowest MSE (best performance) and the models with the highest MSE (worst performance) among the 10 equal size samples for each solvent are shown.....	140
Table 5-5. Details of the logistic regression models taking into account the first three principal components (Model 3) in different solvents. The models with the lowest MSE (best performance) and the models with the highest MSE (worst performance) among the 10 equal size samples for each solvent are shown.....	140

Table 5-6. Single-variable models: details and performance. The models noted as “best” are the ones with the lowest MSE out of the 10 random samples. Similarly, the ones marked as “worst” have the highest MSE. This applies to Table 5-6, Table 5-9 and Table 5-10.	143
Table 5-7. The 10 descriptor combinations with the lowest MSE in the dichloromethane data set – sample 1	145
Table 5-8. The descriptor combinations with the lowest MSE in each of the 10 subsets in the dichloromethane data	146
Table 5-9. Two-variable models: details and performance	147
Table 5-10. Three-variable models: details and performance.....	151
Table 5-11. Relative standard deviation of the intercept, first coefficient and second coefficient over 10 models in each solvent	155
Table 5-12. Average cut-off point in each solvent's dataset	166
Table 5-13. The intercept value in normal and intercept-adjusted models	175
Table 5-14. Part of the AVS_H2 correlation matrix	176
Table 5-15. The average performance of the intercept-adjusted alternative models (over 10 samples) in each solvent's dataset.	180
Table 6-1. Additional examples on the importance of size and branching in solvate formation	194
Table 6-2. Additional paired examples indicating the importance of hydrogen bonding ability in solvate formation. Each pair of examples has similar background shading.....	200
Table 6-3. Accessibility of hydrogen bond donors.....	210
Table 6-4. Additional examples of mispredicted data, involving a short halogen bond in the structure. The examples are given from the chloroform dataset	217
Table 6-5. Additional examples on mispredictions due to the presence of a zwitterion	221
Table 7-1. The molecular structure, some selected molecular descriptors, application and selection strategy of the 10 drug candidates	243
Table 7-2. Prediction vs. screening results for the ethanol predictive model	270
Table 7-3. Prediction vs. screening results for the methanol predictive model.....	271
Table 7-4. Prediction vs. screening results for the dichloromethane predictive model	272
Table 7-5. Prediction vs. screening results for the chloroform predictive model	273
Table 7-6. Prediction vs. screening results for the water predictive model	274
Table 7-7. Crystallographic parameters of the new dichloromethane solvate (II) versus the reported one (I).....	278
Table 8-1. Crystallographic parameters of fenofibrate polymorphs	298
Table 8-2. Intramolecular angles and short contact distances in fenofibrate. The RMSD is a function in mercury program that compares structural similarity between molecules/forms.	301

Table 8-3. Relevant short H-acceptor interactions in the crystal structures of fenofibrate forms I, IIa and III. The H atoms have the same numbers as C atoms that they attached to. (See Figure 8-1 for atom numbering). The symbols A, H and D in the table are short of Acceptor, Hydrogen and Donor atoms 309

Chapter 1: Introduction

Chemical information is growing at exponential rates, as shown by surveys of databases. This provides an easy-to-access, reliable source of data. Investigation of such data has the potential to help in finding new trends, leading to establishing new relationships that were not previously known. The possibility to conduct such investigations has been long recognized, where one of the first examples to obtain knowledge from chemical data took place in 1868.¹ Today, the area of chemical information investigation, also known as Cheminformatics, is a field that is growing, where the number of publications about chemical information has increased lately.² One area in which cheminformatics research can be applied is solid state chemistry, where it is possible to anticipate the physicochemical properties of materials using previously obtained knowledge. The information obtained from X-ray diffraction could be particularly useful in solid state chemistry. This is due to the large amount of information available. Gavezzotti in 1998 stated that "X-ray crystallography is even today a potential source of a wealth of physicochemical information which awaits to be tapped."³ However, with increased complexity of the problems and extent of the data available, such studies become harder. Nowadays, cheminformatics research is a multi-disciplinary area in which mathematical, statistical, programming, theoretical and experimental chemistry knowledge is required. It is important to mention that the development that is seen in all these fields has also facilitated conducting cheminformatics research.

In solid state chemistry, a phenomenon that is poorly understood today is solvate formation. Solvate is a crystalline solid form in which two or more materials constitute the crystal structure, where at least one of these materials is in the liquid state at room temperature. This form has received special attention in the past few years due to its potential beneficial implications, where it could be used to obtain desired properties in a solid. It also has the potential to have harmful implications, where its unexpected formation could lead to a change in the physicochemical properties of the manufactured material. Additionally, the presence of organic solvents as part of the crystal structure could cause toxicity. Despite the large amount

of data available on solvate forms, where around one third of organic chemical crystalline materials form hydrates,⁴ the reasons for solvate or hydrate formation are not explicitly known. Trial-and-error approaches are nowadays used (in industry for example) in order to rule out the ability of a material to form a solvate. In this work, we use a knowledge-based approach in an attempt to relate the chemical features of organic compounds to solvate formation. This could reveal a connection between the molecular structure and the ability of a material to form a solvate. This is achieved by the information from single-crystal X-ray diffraction experiments recorded in the Cambridge Structural Database.

In this thesis, Chapter 2 gives a literature background regarding the areas that are relevant to this work, including solid-state chemistry, statistics and experimental methods. The materials and the methods used are summarized in Chapter 3. Chapter 4 presents a detailed report on how the data of this project was collected from the Cambridge Structural Database, while Chapter 5 focuses on the data mining and statistical modelling of the collected data. It also shows detailed analysis and criticism of the resulting predictive models. Chapter 6 gives a closer look on the factors included in the predictive models and the factors that could probably be added to improve the models via examples.

The applicability of the predictive models in the pharmaceutical industry is tested in Chapter 7, where predictions obtained by the models are compared to the results obtained by experimental work. This chapter also reports PXRD patterns of new solvate forms and a detailed single crystal structural report of a griseofulvin dichloromethane solvate.

Chapter 8 shows results of a collaborative project with a department colleague, Mr. Pratchaya Tipduangta, who studied the heterogeneous crystallization of fenofibrate under different conditions. The chapter focuses on the crystal structures of fenofibrate forms IIa and a newly found form III.

1.1 References

1. Brown AC, Fraser TR. On the Connection between Chemical Constitution and Physiological Action; with special reference to the Physiological Action of the Salts of the Ammonium Bases derived from Strychnia, Brucia, Thebaia, Codeia, Morphia, and Nicotia. *Journal of Anatomy and Physiology*. 1868;2(2):224-42.
2. Nantasenamat C, Isarankura-Na-Ayudhya C, Naenna T, Prachayasittikul V. A PRACTICAL Overview of Qantitative Structure-Activity Relationship EXCLI Journal. 2009;8:74-88.
3. Gavezzotti A. The Crystal Packing of Organic Molecules: Challenge and Fascination Below 1000 Da. 1998. p. 5-121.
4. Clarke HD, Arora KK, Bass H, Kavuru P, Ong TT, Pujari T, et al. Structure–Stability Relationships in Cocrystal Hydrates: Does the Promiscuity of Water Make Crystalline Hydrates the Nemesis of Crystal Engineering? *Crystal Growth & Design*. 2010;10(5):2152-2167.

Chapter 2: Literature review

2.1 Solid forms of pharmaceuticals

Pharmaceutical ingredients are often produced in the solid form. However, solid phases can be either single or multi-component. There is a great variety in the possible compositions of multi-component organic solids and this situation has caused discussions about feasible classification systems of organic solids.^{1,2} According to the most recently proposed classification systems, three main classes of multicomponent solids are distinguished. These are salts, cocrystals and solvates.² Clear definitions of these classes are not available and are often interpreted in various ways – this will be further discussed in the following sections. The classification systems referred to, however, generally recognize that a salt contains two ions, a cocrystal contains two neutral molecules that are solids in their pure form and solvates contain neutral molecules one of which is liquid in its pure form at ambient conditions. It is possible that a multicomponent solid contains more than two components and therefore subclasses are formed. For example, a material containing two ions and a neutral molecule is a cocrystal salt and a solid containing three molecules one of which would be a liquid in its pure form is a solvate of a cocrystal.²

Each of the single- or multi-component solids discussed can be either crystalline or amorphous: crystalline solids exhibit long range structural order while amorphous materials are structurally disordered. Crystalline materials, both single- and multi-component, potentially can show polymorphism – different crystal structures with the same composition.

Research on solid forms of pharmaceutical compounds is of interest for pharmaceutical companies because of several reasons. Firstly, each solid form has unique physicochemical properties and therefore needs to be studied separately. Secondly, each solid form is also subject to intellectual property protection.³

2.1.1 Amorphous materials

The significance of amorphous materials in pharmaceutical solid-state chemistry is related to two aspects. Firstly, although the stability of an amorphous form is inferior to that of a crystalline form, it is still sometimes selected for production because of better solubility and dissolution.⁴⁻⁶ Different techniques are employed by the pharmaceutical industry to stabilize amorphous materials. Some examples of medicines containing amorphous API are Lopinavir,^{7,8} Cefuroxime axetil^{6,9,10} and Zafirlukast.^{6,11} Secondly, amorphous form can be an intermediate phase of solid-state reactions.⁴ Amorphous materials can be either single-component or multicomponent materials.

Although amorphous materials can have some short-range order (for example, interactions to neighbouring molecules), translation or rotational order cannot be identified in these solids.^{4,12} This means that there is no three-dimensional long-range order in amorphous materials and therefore they are similar to liquids.¹³ The main tool used to distinguish between amorphous and crystalline structures is X-ray diffraction.¹³ However, it has been noted that a clear definition of amorphous does not exist, due to a continuum between amorphous and crystalline states.⁴ Ideal crystalline materials are rare, and real crystals usually show some degree of disorder and presence of crystal defects. Furthermore, it is known that the particle size can influence properties of a material. For example, a crystalline material with very small (Nano sized) particles would behave similarly to amorphous material and care must be taken to distinguish between these states.⁴

Common methods for preparation of amorphous solids include quench cooling of a melt, freeze-drying (lyophilisation), spray-drying and mechanochemical treatment (milling).^{14,15} Amorphisation can also be achieved by fast precipitation and by desolvation of solvates.¹⁵ In the latter case, desolvation of a solvate leads to disintegration of a crystal lattice forming an amorphous material.

2.1.2 Polymorphism

Polymorphism is an ability of a compound to crystallize in different arrangements.¹⁶ Two polymorphs can differ in the conformation of the molecules (conformational polymorphism), arrangement of the molecules in space (packing polymorphism) although most commonly both these situations are encountered at the same time. Polymorphic forms of a compound therefore have different crystal structures and consequently – different stability and physicochemical properties.¹⁶ This is especially important in the pharmaceutical industry because two polymorphs with different properties (for example, solubility) can change the efficacy of the drug and lead to inconsistency of dosing. Because of this, polymorphic purity of a drug compound is highly important and pharmaceutical companies need to ensure control of polymorphism of their API's. Control of polymorphism is achieved *via* different methods, such as seeding, use of additives, templating, solvent control, use of membranes, confined space.¹⁷

Crystalline phases are always more stable than amorphous phases, however, the stability of different polymorphic forms also can vary in a wide range. In order to explore the possible polymorphic forms of a compound, crystal energy landscapes can be calculated.¹⁸ At given conditions one of the polymorphic forms of a compound will be more stable than other forms - this form has the lowest free energy with regard to other crystalline forms. All other polymorphs are metastable. Every two polymorphic forms can be either monotropically or enantiotropically related.^{16,19} In a monotropic system one of the two polymorphs is more stable at temperatures up to its melting point while in enantiotropic system a transition point exists. This means that under the transition temperature one of the polymorphic forms is more stable than the other and after this point the other form becomes more stable, while at the transition temperature both polymorphic forms have equal stability.^{20, 21} Graphically, this can be illustrated by the Gibbs free energy curves as shown in Figure 2-1.

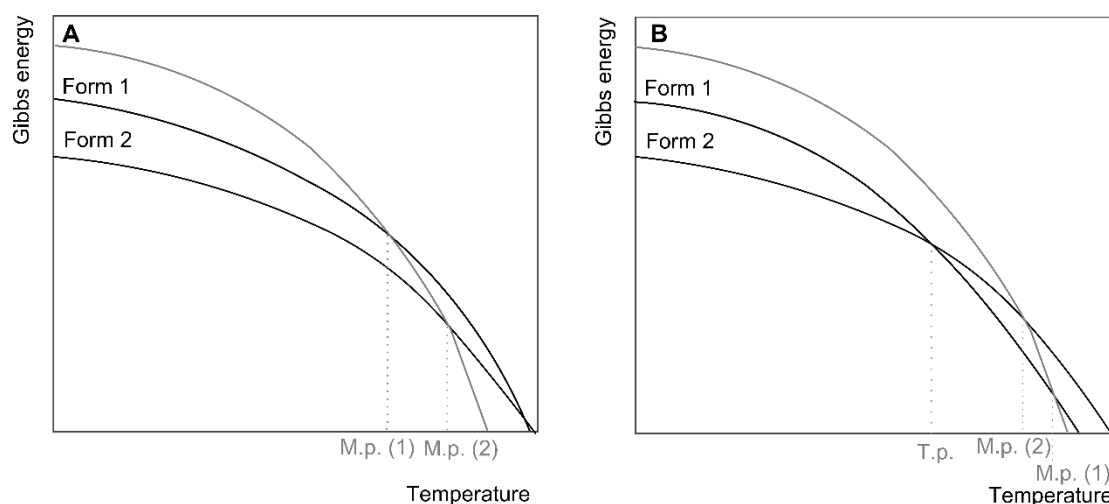


Figure 2-1. Graphical illustration of Gibbs energy in monotropic polymorph system (A) and in enantiotropic polymorph system (B). M.p. is melting point of Form 1 or Form 2 correspondingly and T.p. is polymorph transition point in enantiotropic system.²²

It is important to distinguish between monotropic and enantiotropic polymorph systems in order to prevent undesired polymorph transitions in pharmaceutical products. This is because in a monotropic system solid-solid phase transitions are irreversible, while in an enantiotropic polymorph system reversible transitions can take place.

In order to distinguish between monotropic and enantiotropic polymorph systems Burger and Lamberger have established several rules.^{20,23} The Heat of Fusion rule states that “If the higher melting form has the lower heat of fusion the two forms are usually enantiotropic, otherwise they are monotropic”.²⁰ The heat of transition rule states that “if an endothermal transition is observed at some temperature, the two forms are related enantiotropically. If an exothermal transition is observed, the two forms are either related monotropically or the transition temperature is higher”.²⁰ Additionally, it has also been pointed out that a structure with lower density will be less stable at absolute zero and that analysis of absorption bands in IR spectra also allows to compare stability of two forms.²⁰

2.1.3 Multicomponent crystalline solids

Alongside the single-component organic solids, multicomponent solids, such as solvates, cocrystals and salts, are also possible.

2.1.3.1 *Cocrystals*

According to FDA definition cocrystals are “Crystalline materials composed of two or more different molecules within the same crystal lattice that are associated by nonionic and noncovalent bonds”.^{1,24} The scientific literature is not consistent when discussing whether solvates (and hydrates) should be classified as cocrystals. It has been pointed out that from supramolecular perspective solvates and cocrystals are related.¹ From practical perspective, however, it is beneficial to differentiate between cocrystals and solvates.²⁵ According to Aakeroy,²⁶ reactants making up cocrystals should be solids at ambient conditions. In his perspective, cocrystals are structurally homogeneous crystalline materials containing neutral molecules in stoichiometric amounts.²⁶ According to these rules, solvates, clathrates and salts cannot be seen as cocrystals. Although, the rules set out by Aakeroy²⁶ seem to make a distinction between cocrystals and salts, a continuum exists between these species too as partial transfer of proton is possible in molecular complexes.²⁷

Cocrystals are commonly obtained based on knowledge of hydrogen and halogen bonds between functional groups of the reacting molecules.^{28,29,30} Other interactions, such as van der Waals forces and π - π interactions can play a role cocrystal formation.³¹ The necessary information can be obtained by analysing the available crystal structure data of other cocrystals. Several studies employing Cambridge Structural Database (CSD)³² have shown the prevalence of certain supramolecular synthons.^{33,34,28} This information can be used to design new cocrystals.

Until recently, the production of pharmaceutical cocrystals was restricted due to FDA guidelines. Currently (since 2013) cocrystals have been recognized as drug intermediates and the industrial interest about them has increased.

2.1.3.2 *Solvates*

In comparison to cocrystals, in solvates one of the constituents is liquid at ambient conditions.²⁵ Solvates of a pharmaceutical compound are sometimes referred to as pseudopolymorphs or solvatomorphs although recently the term “solvates” has been preferred.²⁵ Solvates can contain either a stoichiometric or a non-stoichiometric amount of solvent. These two types of solvates are being referred to as stoichiometric and non-stoichiometric solvates.^{35, 36}

In a stoichiometric solvate the solvent molecules are usually a crucial part of the crystal structure, being bound to the drug molecule by specific intermolecular interactions. Incorporation of a solvent molecule in such a structure leads to a more stable crystalline form therefore stoichiometric solvates can be chosen as a final drug product to be marketed.^{37, 38} In non-stoichiometric solvates, on the other hand, the solvent molecules usually do not form strong interactions to the host molecules.

An alternative way to classify solvates (this classification system is typically used for hydrates) is with regard to their structure: isolated site solvates, channel-type solvates and ion-associated solvates can be distinguished.³⁹ In isolated-site solvates, solvent (water) molecules are separated from each other and connected to API molecules by hydrogen bonds. Isolated-site solvates are usually stoichiometric. Some examples of stoichiometric isolated-site solvates are cephadrine dihydrate, cefaclor dehydrate⁴⁰ and siramesine hydrochloride monohydrate.⁴¹

In channel-type solvates the solvent (water) molecules are accommodated in channels or between layers of the host molecules (API)³⁷ (these can be both stoichiometric and non-stoichiometric). Channel and layer type structures can often accommodate a range of various solvents resulting in series of isostructural solvates. These solvates are often non-stoichiometric and easily undergo desolvation. In case of isostructural solvates, desolvation commonly leads to isostructural (isomorphic) desolvates with the same structure of the host molecules and empty voids that used to accommodate solvent molecules.⁴² Some compounds known to form isostructural solvates are tenofovir disoproxil fumarate⁴³ and sulfathiazole.⁴⁴

In ion-associated solvates solvent/water molecules are coordinated around an ion. This type of solvate is not the scope of this study, as we focus on organic solvates.

It is possible for a solvate (with the same stoichiometry) to have several crystalline forms – polymorphs.

Manufacturing solvated forms of API's is limited by the toxicity of the solvent present in the material. Hydrates, that is solvates of water, however, are free from this concern and therefore are the most commonly used solvates. Additionally, the water molecule because of its small size can easily fill structural voids. Moreover, its hydrogen bond donor and acceptor properties ensure efficient bonding to API molecules resulting in a stable crystal lattice.³⁶ It has been estimated that at least one third of pharmaceutical compounds can form hydrates.⁴⁵ Furthermore, water is present in the atmosphere and can come into contact with the drug compound during processing of the solid.⁴⁶ Many hydrate forms have been commercialized, for example, amoxicillin trihydrate,⁴⁷ darunavir ethanolate⁴⁸ and dasatinib monohydrate.⁴⁹ Solvates containing multiple solvents are much more rare among the manufactured products. An example on the latter is indinavir sulfate ethanolate.⁵⁰

While cocrystals and salts are usually prepared intentionally, solvates can form unexpectedly. Since it is important to obtain a pure solid form as a final product of manufacturing, information on possible solvation of a compound is crucial.

2.2 Identification of the solid form

Most of newly discovered Active Pharmaceutical Ingredients (API) do not reach the market, where it is reported that 89 % of these APIs fail before being marketed.⁵¹ The main reasons for this failure is their poor biopharmaceutical properties, low efficacy and toxicity, where these account for 41 %, 31 % and 22 % of total failure in development, respectively.⁵² Therefore, the sooner the properties of an API are known, the less time and money is spent on their development. In order to achieve this early judgement of which medical candidates are going to be useful, two main approaches are available, these are an in vitro (experimental) and in silico (theoretical) approaches, which are going to be discussed in sections 2.2.1 and 2.2.2.

2.2.1 Experimental identification of solid form

Historically, screening experiments in biological and pharmaceutical areas were slow processes that used to consume a considerable amount of materials.⁵³ A report in 2007 was published by Pereira and William who worked in Pfizer. This report mentions that each biological or pharmaceutical assay required to be 1ml in volume, required an amount of the compounds being investigated and a separate test tube for each experiment.⁵⁴ This resulted in testing 20-50 compounds in a laboratory per week. In the past two decades, it was possible to automate and speed up the process using computers and robotics. Currently, it is possible to use picolitres of liquid and conduct over 100,000 assays per day *via* Ultra High-throughput screening (uHTS), resulting in reduced time and expenses of a study.^{55,56} Note that the process of screening can be used in various areas such as conducting biological, toxicological and pharmacokinetic assays.⁵⁵

In this work, we highlight the application of high-throughput techniques in solid form selection. A solid form is normally obtained by crystallization, as the final step of synthesis. Multiple crystallization methods are known, such as evaporative, cooling and anti-solvent crystallizations. For the same compound, using a different crystallization method can result in a different solid form. Additionally, in any of these crystallization processes it is possible to change the form of the solid obtained by manipulating factors such as the composition of the crystallization medium, concentration and additives. These factors are summarized in a review that was published by Sherry L. Morissette *et al* in 2004.⁵³ The variety of methods of crystallization and the number of factors to adjust in each result in a large number of experiments that need to be conducted to cover the possible solid forms of an API. For this reason, high-throughput has become one of the standard techniques that companies use for discovering solid form diversity. Although this is a popular method nowadays, it still has its disadvantages when applied in solid state chemistry. For example, it can require large amounts of material, especially when crystallizations are conducted. This is because crystallization experiments cannot be done at the pico-scale these methods offer. Additionally, if experiments did not cover absolutely all possible materials/ratios of materials in a crystallization experiment, unexpected solid forms could arise.⁵⁷

2.2.2 Theoretical prediction of possible solid forms

2.2.2.1 *Crystal Structure Prediction (CSP)*

Crystal structure prediction is a computational chemistry method, aiming to find the crystal structure of a molecule given its chemical structure and in some cases, crystallization conditions.⁵⁸

For a specific organic structure, current CSP methods work by looking for the structure with the lowest lattice energy (also known as global minimum) among all possible structural arrangements. This means the found structure is likely to have high thermodynamic stability.

Note that it is not the absolute most stable form because the entropy is ignored here.⁵⁹ The energy of these structures is mainly calculated using empirical models or ab initio calculations or a mixture of both.

Although this might sound straightforward, different challenges are faced; most notably the number of possible structures (largely depends on the space group choice and the number of independent molecules in the space group), the phenomenon of polymorphism and the choice of the model to perform the calculations.^{60,61} For example sometimes the structure that has the lowest lattice energy is not experimentally observed. Sometimes one polymorph is the most stable under certain temperature, but as this temperature changes, a different polymorph becomes more stable.⁶¹

In order to evaluate the different methods available and the progress in improving the predictability of crystal structures, the Cambridge Crystallographic Data Centre (CCDC) periodically organizes international blind tests of crystal structure predictions. In this test, research groups that are interested in developing crystal structure predictive methods are given a set of organic structures with unknown crystal structures. The groups are then asked to report their predictions regarding the space group, cell dimensions and atomic coordinates of the given structures.⁶¹ Six of these tests have been conducted so far, with the latest of them being in late 2015. The details about the first five of these tests can be found in the references provided.^{58, 62-65}

It is worth noting that the fourth CSP blind test has introduced multi-component solids as a new class to be part of the blind test.⁵⁸ This emphasizes the growing importance of multi-components in solids state chemistry. Specifically, predicting the crystal structure of a cocrystal was attempted. Determination of the cocrystal structure was the hardest among all others, as the report of the test states:

“As expected, the cocrystal was the main problem – the increased search space was the main reason for including the new category of two-component crystals in this blind test and seven of the 12 groups who attempted predictions for this system did not locate the observed crystal structure in their search.”⁵⁸

Few years later, multi-components still showed a big challenge to the process of CSP, where the report of the fifth blind test states:

“Hydrate (XXI) proved to be one of the most challenging systems in the blind test”⁶⁵

Most recently, the report on the sixth blind test included a prediction of a crystal structure of a chloride salt hydrate. One of the participating groups was able to predict the experimentally known structure, where they ranked it as the second most stable structure in their submitted prediction. Although this prediction doesn't directly fall into the scope of this thesis, a correct prediction of the experimental structure of a 3-component system including a solvent is important to note. Despite these improvements, some issues remain the limiting factors of CSP. An example is the complexity of this method; where it requires an expert in the field to generate and assess the possible structures. Additionally, it has a high computational cost, where a landscape calculation of a molecule can take months.⁵⁷ Therefore, at current times CSP is used as a complementary method with experimental work and high throughput crystallizations, where it helps highlighting the possibility for undiscovered polymorphic forms.⁵⁹

2.2.2.2 Cheminformatics

Cheminformatics (also known as chemoinformatics, chemical informatics and molecular informatics) can be defined as the science in which information technology is employed in order to help making better and faster decisions in the fields of drug discovery and

development.^{66,67} This is a multi-disciplinary science that employs the fields of chemistry (chemical information) of structures, mathematics, statistics and computer programming.⁶⁶ Chemometrics, a sub-field of cheminformatics is a field that is concerned with deriving chemical information from experimental data.⁶⁸ In biology, two parallel sciences that are heavily used nowadays are Bioinformatics and Biometrics. As the name implies, they are similar to cheminformatics but they use data derived from biology.⁶⁹

One of the main concepts in cheminformatics is Quantitative Structure-Activity Relationship (QSAR). This term describes a process in which the structure of a chemical is mapped to the biological activity. A closely related term is Quantitative Structure-Property Relationship (QSPR). This is a process that tries to find an association between the structure of a chemical and a physicochemical property of this structure.⁷⁰

The basis of this field could probably be assigned to Crum-Brown in 1861, where he suggested the possibility of finding mathematical model that would explain chemical theories.⁷¹ He was able to reveal the association between the water solubility of primary alcohols and their toxicity.⁷⁰ Few years after that, he supported this opinion in a publication with Fraser, where their publication focused on the possibility of linking a physicochemical property of a chemical to its physiological action of a chemical.⁷² Several studies on QSAR/QSPR followed after that, as given by review articles.⁷⁰ Currently, knowledge-based cheminformatics approaches play a significant role in lead discovery and optimization, where it assesses the toxicity, permeability and other properties of drug candidates. It is also used in early selection of drug candidates that are going to be tested experimentally using high-throughput techniques.^{73,51}

In order to establish a quantitative structure-activity or structure-property relationship, a few elements are required. Firstly, it is important to have a reliable source of data regarding the chemical structures from which information and later on, knowledge can be obtained. In this

work, we obtain our data from the Cambridge Structural Database (CSD).³² Secondly, a numerical description of the chemical structures is required.⁷⁰ Such description is often known as molecular descriptors (sometimes referred to as “descriptors” alone). More formally, molecular descriptors have been defined by Todeschini and Consonni to be *“The final result of a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment.”*⁷⁴ In this thesis, the molecular descriptors are obtained using a software called Dragon.⁷⁵ The details about the descriptors that are calculated by this software are given in section 3.2.1.

The third step in finding a QSAR/QSPR would be the conduction of statistical tests and machine learning methods in order to fit models that can best describe the data provided and make predictions of future data points. It is also important to evaluate these models to ensure they are not biased.⁷⁰ The statistical part is going to be discussed in details in section 2.3.

2.2.3 Previous CSD investigations on hydrate and solvate formation

Considering the significance of solvates, especially in pharmaceutical science, several attempts to develop approaches for predicting solvate formation have been made. Alongside studies that have investigated the behaviour of compounds that readily form large number of solvates,⁷⁶⁻⁷⁸ A number of studies have been conducted using large datasets obtained from CSD.³² CSD offers valuable information of crystal structures and therefore allows studying various structural aspects that are expected to facilitate hydrate formation.

It was noticed, however, that most if not all of the studies conducted were at least partially based on the consideration that formation of hydrogen bonds are the main governing force of solvate (hydrate) formation. For example, a study conducted by Desiraju investigated the formation of hydrates in relation to hydrogen acceptor and donor properties.¹⁴³ The inspection

of hydrate structures of organic compounds revealed that almost all of these compounds did contain hydrogen bond donor and acceptor groups. The results of this work implied that most of hydrate-forming compounds have larger number of hydrogen acceptor groups than hydrogen donor groups. Regardless, actual hydrogen bonding between the compound and water did not always take place. In some cases (no estimate given) the water only acted as space-filler.

The conclusion made by Desiraju regarding the effect of hydrogen donor/acceptor ratio was countered by another study based on CSD data, performed by Infantes *et al*⁷⁹ who investigated hydrate formation in relation to the count of hydrogen bond donors and acceptors in the compound. In this work, several parameters describing formation of donor-acceptor bonds were derived and calculated for molecules of interest. A number of molecular parameters, such as atom count and dipole moment were also calculated. This study concluded that donor/acceptor ratio does not have effect on probability of hydrate formation. However, this work showed that higher sum of all donor and acceptor groups in the compound facilitate hydrate formation. More polar surface of molecule also was found to facilitate hydrate formation.

Another study, performed by Nangia and Desiraju,⁸⁰ inspected solvates with the 10 most common solvents in the database. The results showed that solvates are more often formed with solvents (1,4-dioxane, DMSO and DMF) that have high probability to participate in multi-point hydrogen bonded recognition schemes between solvent and solute molecules. Hydrogen bonding between solvent molecules was also common.⁸⁰ On the other hand, solvents that have comparatively poor multi-point hydrogen bonding ability (ethanol, ethyl acetate and diethyl ether) were found to rarely form solvates (the occurrence of solvates was calculated considering how often the solvent is used for recrystallization). They also recognized that

solvents such as benzene, p-xylene and CCl₄ are included mostly in rigid framework-type structures as guest molecules.

The crystallographic information available in the CSD has been used to thoroughly characterize the molecular environment of the two most common solvent molecules found in solvate crystal structures – water⁸¹ and methanol.⁸² It was found that the most common environment for water is such that it allows formation of three hydrogen bonds – two involving its hydrogen atoms and one with its oxygen atom. No correlations between the environment and the hydrogen bond strength were found. The inspection of the environment of methanol molecules in solvate crystal structures showed that in 305 of 375 methanol solvate examples one of five molecular environments could be identified. In the most common environment, the hydroxyl oxygen acted both as hydrogen donor and acceptor.

A significant contribution to analysis of solvate crystal structure data has been given by van de Streek and Moteherwell⁸³ who developed a software for CSD data analysis to find sets of structures containing both solvated and non-solvated forms of any compound. They used this in-house software to analyse such aspects as packing density, flexibility of the compound, chirality, number of possible donors and acceptors etc. The study showed that larger molecules commonly include larger number of water molecules (higher stoichiometry). They also found that certain groups (R₂PO₂⁻, Cl⁻ and NH₃⁺) are considerably more common in hydrates than in anhydrides. Some other groups (CF₃, CCl and OCONH), however, were preferred in anhydrides. A comparison of contacts formed by the same compound in hydrated and anhydrate structures showed that for most of the functional groups the count is higher in hydrates. This was most obvious for Cl⁻, COO⁻ and NH₃R⁺ groups. This study did not find correlation between flexibility of molecules and tendency to form hydrates. Chirality, on the other hand, had a positive effect on hydrate formation. In a subsequent study, van de Streek developed a program that screens the CSD structures to find solvates based on topological

indices.⁸⁴ This program allowed obtaining lists of solvates for compounds with the 51 most common solvents. Closer inspection of the extracted lists showed that promising solvate formers are molecules with non-coplanar aromatic rings, cholic acid derivatives and calixarene/cyclodextrin type molecules. The solvate lists generated by this program allow preparing a correlation matrix that can be useful to predict solvation of a compound in certain solvents based on knowledge about similar cases.

The reports reviewed in this section show that although invaluable work has already been performed to understand the structural reasons of solvate formation, more investigations are needed in this area. For example, although it has been extensively shown that intermolecular interactions, especially hydrogen bonds often facilitate solvate formation, some significant factors such as flexibility of molecules, their size, steric hindering and other have not been properly challenged. Moreover, it is expected that most of organic structures present in CSD would contain heteroatoms, therefore presence of hydrogen bond donors and acceptors is unavoidable, and it is not surprising to find that most of the compounds forming solvates would also have some hydrogen bond donors and acceptors. Regardless, only a fraction of all heteroatomic molecules actually does form solvates. Therefore, it would be useful identify the factors governing solvate formation and quantify them. This would provide a tool that would be able to predict the formation of solvate. In order to develop a comprehensive approach that would be able to predict solvate formation, as many aspects as possible should be objectively evaluated.

2.3 Statistics

In this section, statistical methods that were used in the thesis are going to be presented.

2.3.1 Hypothesis testing

In its simplest form, this process compares two contradicting hypotheses about a data set, in terms of a variable. The result of the test is the decision of accepting one hypothesis and rejecting the other. This decision is based on the value of a statistic that is calculated through the values of the variables of the data set.⁸⁵ In this thesis, a comparison between two distributions, whether they come from the same population or not is required. the Wilcoxon signed-rank test.⁸⁶ The choice of this test was based on its properties, where it does not assume a normal distribution of the data (non-parametric).⁸⁷ Additionally, this test is ordinal which means it is not going to be affected largely by outliers.⁸⁸ Such properties are suitable for the type of data in the problem being solved in this thesis. The study being conducted is for thousands of molecules, a normal distribution cannot be guaranteed and the amount of outliers will differ depending how each molecule is described.

This test, developed in 1947, works by comparing two samples in order to know if they come from the same population. The test is closely related to the Wilcoxon rank-sum test which was developed earlier in 1945.⁸⁹ The Wilcoxon signed-rank test works with paired samples, while the Mann Whitney test was designed to work when the two samples being tested are of different sizes.⁸⁶

The null hypothesis (H_0) is that the two samples come from the same population against an alternative hypothesis (H_1) that the two samples come from different populations. One commonly used method for accepting or rejecting the null hypothesis is comparing a chosen alpha (α) level to the p-value obtained by the test.⁹⁰ Alpha is simply the significance level that is chosen by the test conductor. The p-value is a probability of obtaining a statistic value that is as extreme or more extreme to the specified boundary value.⁹¹ It is important to note that the p-value is calculated based on the sample being tested, assuming the null hypothesis is true.⁹² The convention is to use a p-value 0.05 as recommended by Fisher.⁹³ When a p-value is below

the significance level, the null hypothesis (H_0) is rejected and the alternative hypothesis (H_1) is accepted. It is worth noting that the alpha value can be on the positive and/or the negative end of the distribution giving three types of test, a left-tailed, a right-tailed and a two-tailed test, as illustrated by Figure 2-2.

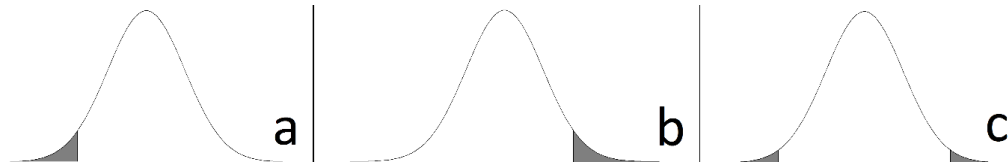


Figure 2-2. (a) A distribution with a left-tailed alpha value and an equivalent $|p\text{-value}|$. (b) A distribution with a right-tailed alpha value and an equivalent $|p\text{-value}|$. (c) A distribution with a two-tailed alpha value and a $|p\text{-value}|$ equivalent to $\alpha/2$.

The choice of the tailing of the test is set by the user. It is important to notice that in one tailed tests the absolute values of the alpha and the p remain similar, but when the a two-tailed test is chosen the p-value is split between the positive and the negative end, causing it to be half of the magnitude of the alpha level, as illustrated in Figure 2-2.

2.3.2 Data mining and machine learning

Data mining can be defined as the process of learning from large amount of input data.⁹⁴ Machine learning, a closely related term, is the use of computers in order to learn from these large amounts of data.⁹⁵ The aim of machine learning can be predictive (to know what future data would look like) and/or descriptive (to know more about existing data).⁹⁵

Typically, machine learning algorithms can be classified into two types, these are the supervised methods, which require previously known outcome of the data (labelled data) and the unsupervised methods, which don't require previously labelled data.⁹⁶ More details about these methods are given in sections 2.3.2.1 and 2.3.2.2.

2.3.2.1 *Unsupervised learning: Principal component analysis (PCA)*

As presented earlier, this type of learning works with unlabelled data. One of the most commonly used unsupervised techniques, and the one used for this project is Principal Component Analysis (PCA).⁹⁷ It was introduced by K. Pearson back in 1901.⁹⁷ It works by projecting the data to a new set of dimensions (variables), which are essentially linear transformations of existing dimensions.⁹⁹ The new imaginary variables (known as principal components) maximize the variance in the dataset in terms of the variables provided. The first principal component accounts for the largest possible variance, and each succeeding variable accounts for the highest possible variance, on the condition that each new of these principal component has to be orthogonal (uncorrelated) to the rest of the principal components. Such properties make it an ideal way for visualizing the data in a low-dimensional space (for example in 2D or 3D plots).¹⁰⁰

The technique is used for several purposes, but one of the most popular reasons to apply this method is data exploration and dimensionality reduction. Additionally, it is used as a preparatory method before applying other classification techniques.¹⁰¹ In simpler terms, it could be used to summarize the data and omit repetitive information from the dataset.¹⁰⁰ An example of a transformation of data in terms of 4 variables to 2 principal components is shown in Figure 2-3.

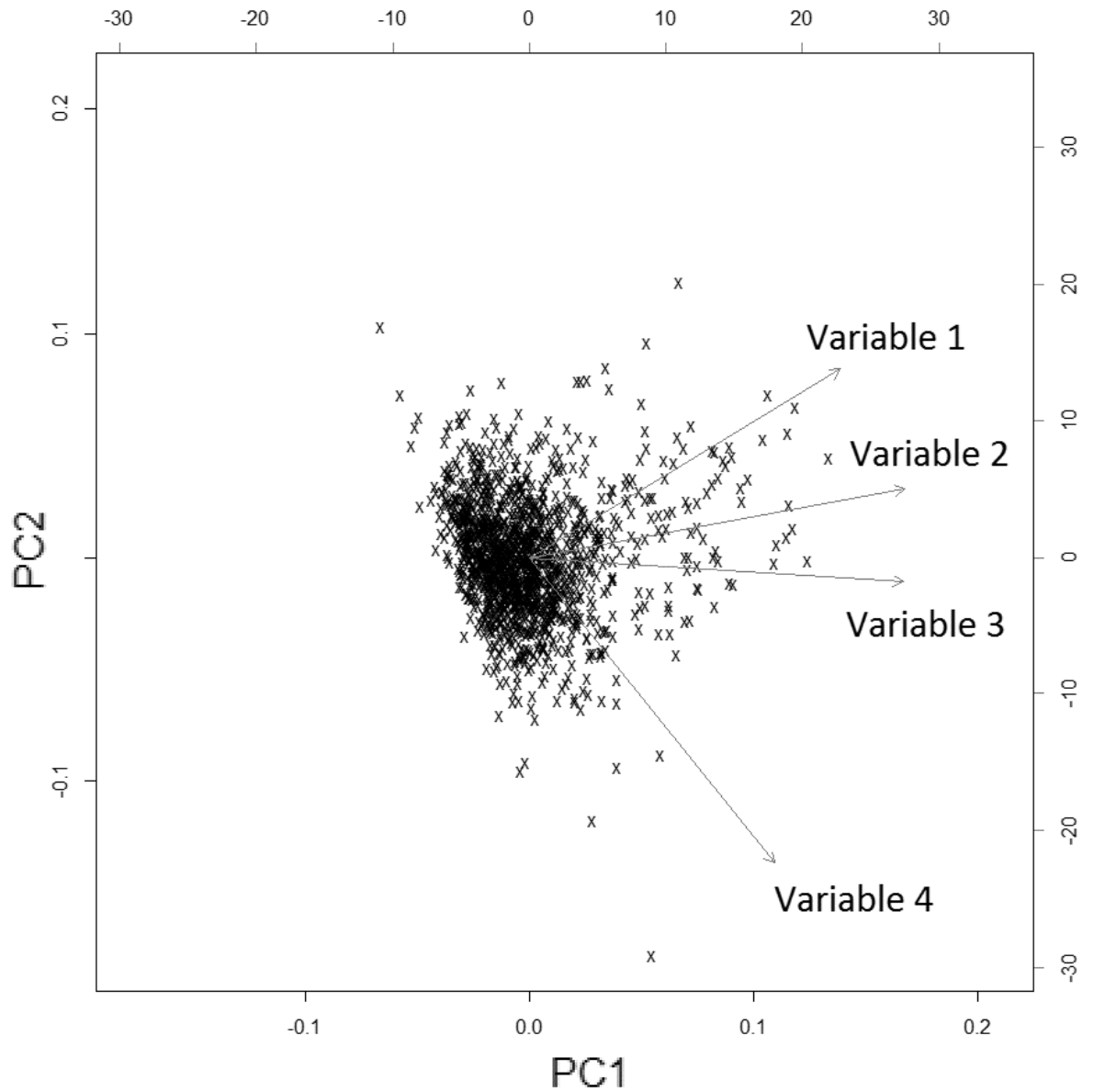


Figure 2-3. An illustration of 4 variables that point in different directions (have some correlation) in terms of principal components 1 and 2.

Principal components can be calculated *via* singular matrix decomposition or by eigenvalue decomposition of the covariance matrix. The software used in this thesis offers both calculation methods but recommends the latter method as for better numerical accuracy in the documentation of the “prcomp()” and “princomp()” functions.¹⁰²

2.3.2.2 *Supervised learning*

In supervised machine learning, we try to find a model that would classify the supplied labelled data correctly in using the predictors (variables) provided in the dataset. The model that is established is then used to make future predictions.¹⁰³ Several supervised machine learning algorithms exist. Some of the popular algorithms are artificial neural networks,¹⁰⁴ logistic regression,¹⁰⁵ and support vector machines.¹⁰⁶ In this thesis, logistic regression, and to a lesser extent, support vector machine are going to be used. An introduction to these methods is provided.

Logistic regression (LR)

To present the idea of logistic regression, a comparison with linear regression could be established. Linear regression is used for finding a relationship between variables. However, in the case of data that belong to two classes (binary outcomes), it is not suitable to fit a linear regression model. For example if the relationship between an independent, continuous variable X and a binary dependent variable Y was to be plotted, the outcome will be data points on two straight lines, with a y value of either 0 or 1, as presented in Figure 2-4.

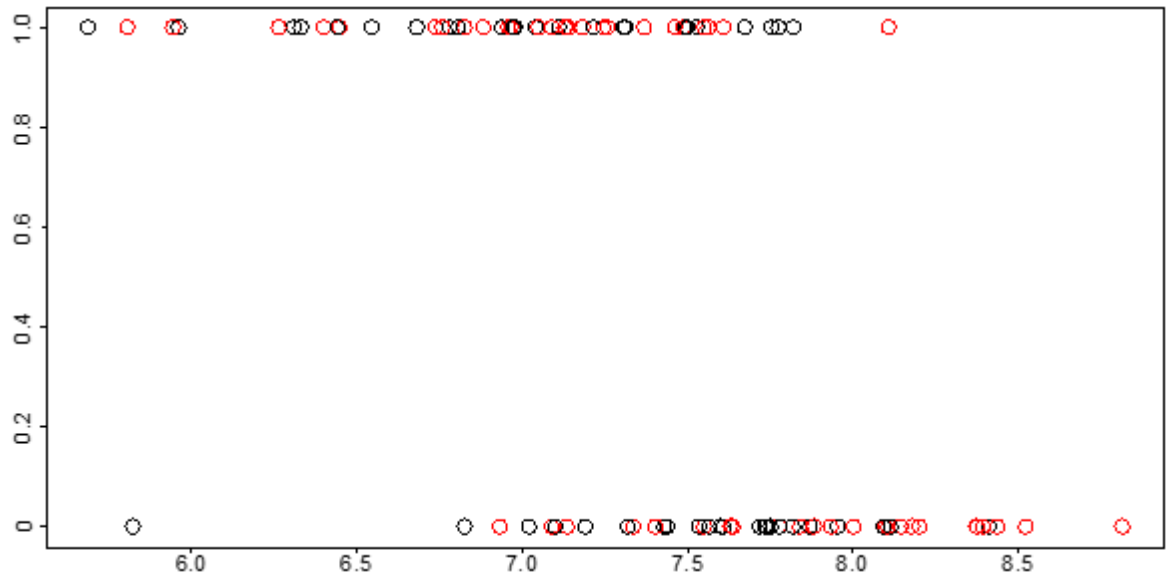


Figure 2-4 Values of Y (on the y-axis) vs the values of the continuous variable X (on the x axis).

It is clear that the relationship between X and Y isn't linear. The logistic regression, a binary classifier that was developed by D. Cox in 1958 proposes a solution to this problem.¹⁰⁷ The idea is to transform the dependent, binary variable Y so that it becomes a linear function of the predictor X.¹⁰⁸ More specifically, it expresses the probability of obtaining the binary response of (Y) depending on the values of the predictor variable (X) using a logistic function as shown in Equation (2-1):

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_i x_i)}} \quad (2-1)$$

where p is the probability of an event to happen, β_0 is the intercept, β_i is the coefficient of the variable x_i . Note that the value of the probability p ranges from 0 to 1 while the value of the term in the exponent ranges from $-\infty$ to ∞ .¹⁰⁹ The relationship between p and x is not linear, plotting their relationship results in a sigmoidal curve as illustrated in Figure 2-5. On the other

hand, this form means that the logarithm of the odds-ratio becomes a linear function of x , as presented in Equation (2-2).

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_i x_i \quad (2-2)$$

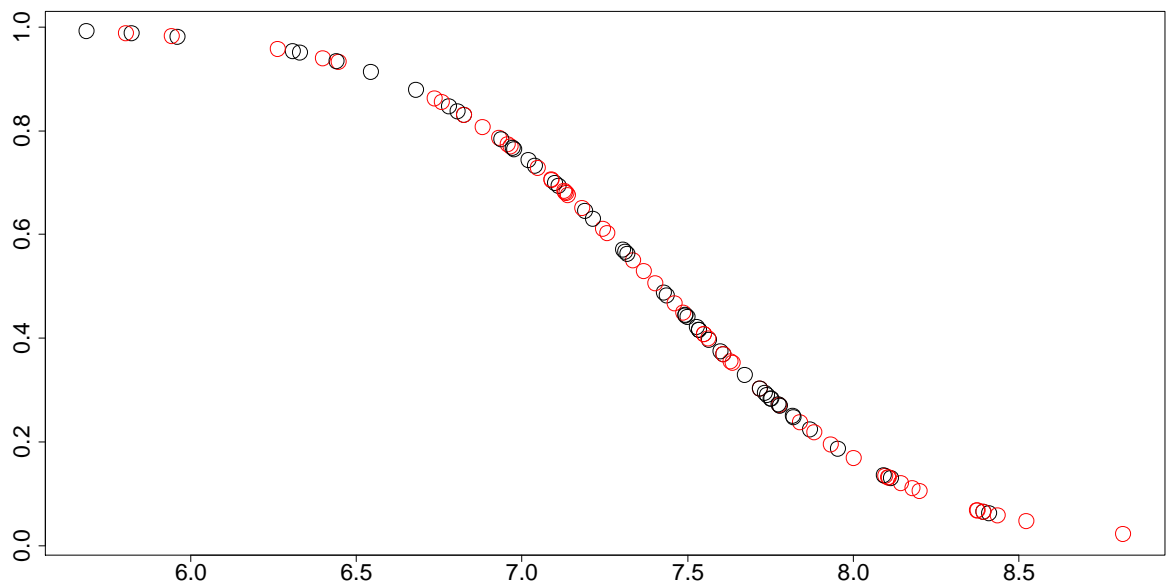


Figure 2-5. Fitted probabilities (Y axis) vs the values of X (X axis). Red and black points correspond to two classes, representing a case of binary data.

The parameters of the function (β, x) in Equation (2-1) are found using a maximum likelihood estimation (MLE) function. Specifically, the software used in this thesis utilizes the iteratively reweighted least squares (IRLS) algorithm, as given by the software documentation.¹⁰²

Support vector machines (SVM)

Support Vector Machines is a supervised machine learning technique that has been introduced by Vapkin in the 1990s.¹⁰⁶ Similar to logistic regression, this algorithm requires labelled data. In its simplest form, the support vector machine algorithm assumes that the given data can be separated linearly in the space of the provided variables (input space),¹¹⁰ as illustrated in Figure 2-6.

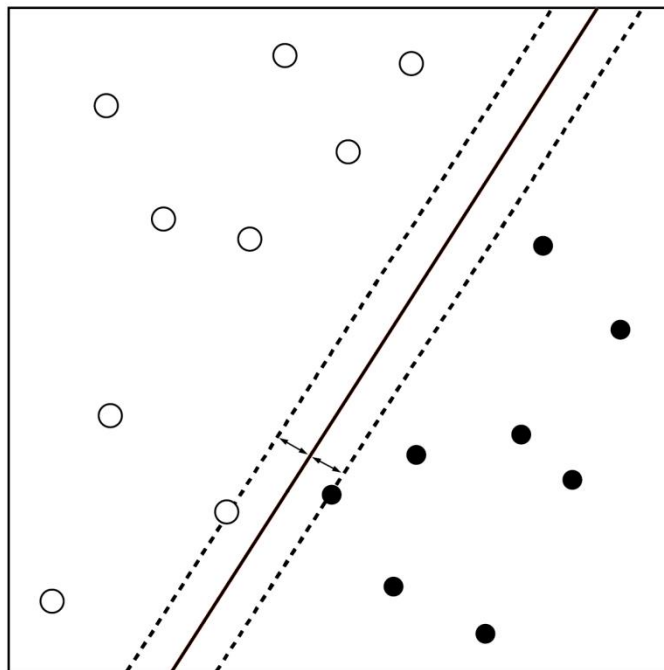


Figure 2-6. An illustration of support vector machine linearly separating binary data.

The hyperplanes are the dotted lines passing through the nearest points from each group; these are known as the support vectors. The middle dotted line is known as the decision boundary.

In support vector machines, the algorithm tries to maximize the margin separating the two classes.¹¹¹ In the case of non-linearly separable data, the same data can be mapped into a

higher dimensional space, known as the feature space F in which the data is linearly separable. Nevertheless, performing this mapping step explicitly could be practically impossible as the dimensions of the new feature space can be infinite in number. Alternatively, a kernel function can be introduced to the SVM algorithm, where it allows us to find a separating hyperplane in feature space F without mapping the data into the feature space.^{112,113} An illustration of non-linear SVM, projected back to the input space is shown in Figure 2-7.

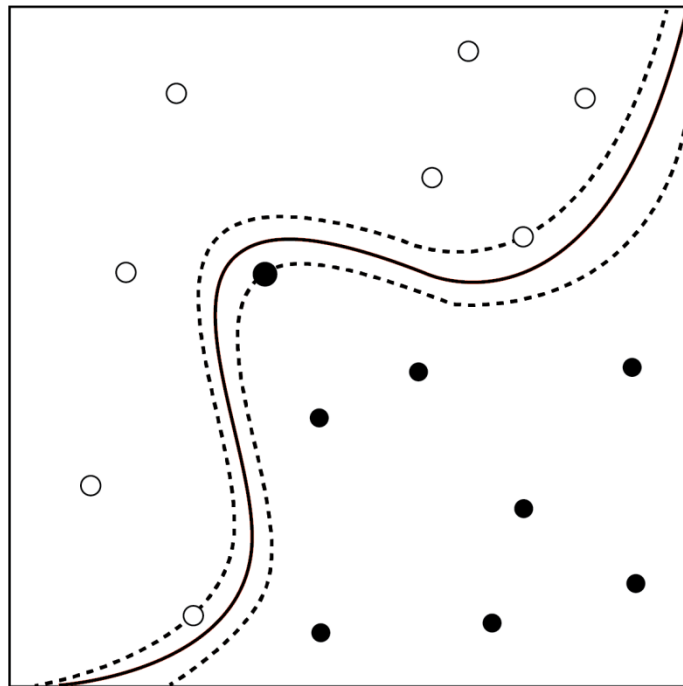


Figure 2-7. An illustration of non-linear vector machine with separating binary data.

Multiple kernels such as the Radial Basis Function (RBF) and the Polynomial Kernel can be used with SVM.¹⁰⁴ The choice of a kernel is normally based on prior knowledge of the data being analysed. Generally, an RBF kernel is used for preliminary testing of SVM.¹⁰⁵ But what if the data is not perfectly separable by SVM? An illustration of such case is shown in Figure 2-8.

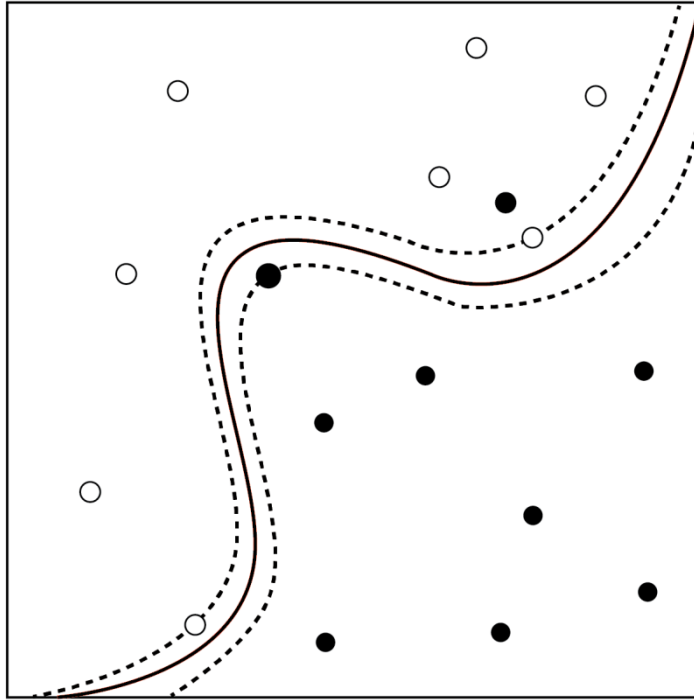


Figure 2-8. An illustration of non-linear support vector machine with soft margins, note that the algorithm converged despite the misclassification of one black point.

In such a case, it is possible to allow a certain number of incorrect classifications by the model, also known as soft margins. This helps increasing the simplicity but reduces the accuracy of the established hypothesis.¹⁰⁶

In comparison to logistic regression, it is possible for SVM to separate non-linear problems, which could be an advantage above logistic regression. On the parameter optimization of Support Vector Machines for binary classification] Additionally, unlike the methods optimized through maximum likelihood estimation, SVM is optimized by structural risk minimization (SRM), which has no prior assumptions regarding the data used (non-parametric).¹¹⁰ This offers SVM another advantage above logistic regression.

2.3.3 Model selection

After machine learning methods are used to fit statistical models, the best performing models need to be selected. The techniques that are used for selecting these are discussed in this section.

2.3.3.1 *Cross-validation (CV)*

This is a statistical method that estimates the error associated with a model trained on a dataset to predict a different, independent dataset.¹¹⁷ Technically, a dataset is randomly split into k number of partitions, where one or more partitions are used for fitting a model and the remaining partitions are used for testing the model. A loop of training and testing is repeated k number of times, until all of the data points have been used for training and for testing the model. Such a procedure helps avoiding overfitting, where the test set is completely different from the training set.¹⁰⁸ An illustrative example of how this method works is given in Figure 2-9.

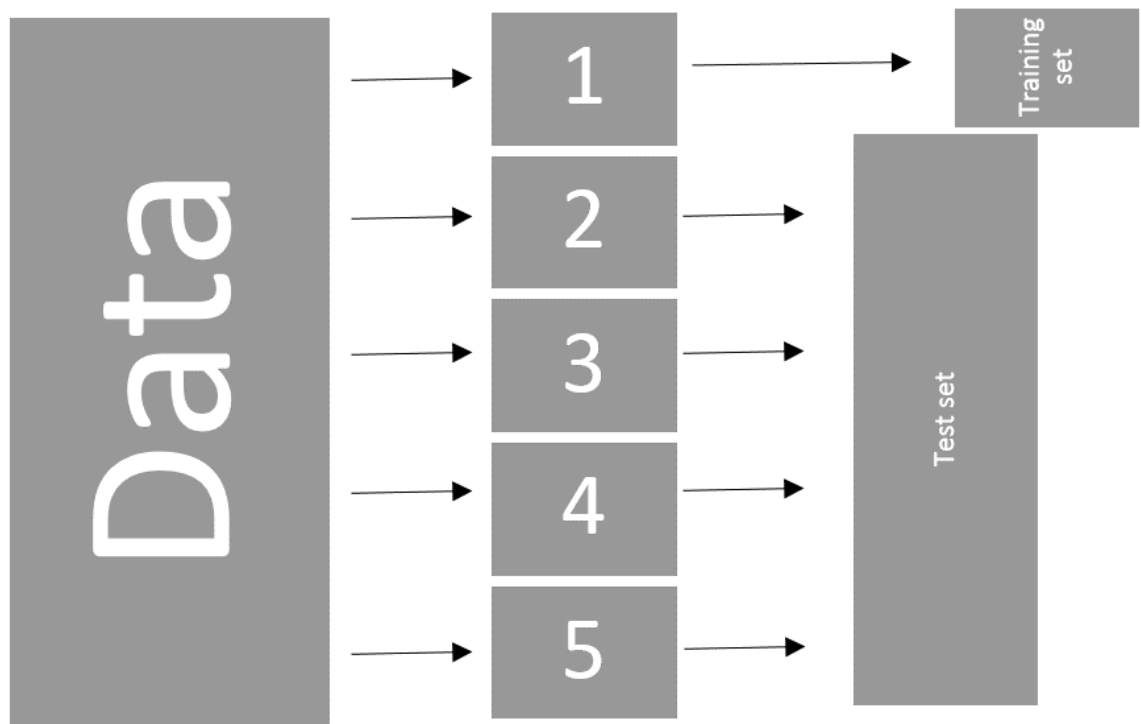


Figure 2-9. An illustration of how a 5-fold cross-validation works.

The fundamentals of Cross validation were introduced by Mosier in 1951.¹¹⁹ The importance of this method in model selection in machine learning has been recognized since the 1970s.¹²⁰ Multiple variations have been applied to this method over time.¹²¹ In this thesis, 10 fold cross-validation is going to be used as a standard method for evaluation of model performance. Although more intense cross validation, such as the leave-one-out method can be performed, researchers argue that it is more reliable to use a moderate number of folds (10-20), not to mention how computationally expensive it is to run a leave-one-out method when the dataset –as it is in the case of this work- is thousands of data points.¹²²

2.3.3.2 Mean squared error (MSE)

Another method for model selection is MSE. This is simply the mean value of the squared loss error of each prediction made by a model. This error estimator takes into account the variance and the bias terms in its calculation, leading to a precise model.¹²³ In this thesis, MSE was the

principal method for model evaluation, where it was obtained for each cross-validation loop, weighted by the sample size of each fold. This is calculated using the formula shown in Equation (2-3):

$$MSE = \frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2 \quad (2-3)$$

where \hat{y}_i is the estimated value from the model, y_i is the real value (0 for solvates or 1 for non-solvates) and n is the number of data points. The average weighted MSE of the 10-fold cross-validation can be calculated using Equation (2-4):

$$\text{average weighted MSE} = \sum_{k=1}^n \frac{N_k}{N} MSE_k \quad (2-4)$$

where n is the number of folds (10), N_k is the sample size in the k th fold, N is the total number of molecules, and MSE_k is the MSE value of the k th fold. Due to the large sample sizes used in the analysis, the weighting has minimal effect on the results. The factor $\frac{N_k}{N}$ will have a value very close to 0.1 for each fold, even if the number of molecules is not divisible by 10. For simplicity, the average weighted MSE calculated by the software is going to be referred to as MSE from this point onward.

2.3.3.3 *The area under the curve (AUC)*

The area under the Receiver operating characteristic (ROC) curve, (see section 2.3.4.1 for more about ROC curve) also known as AUC is a popular estimate that can be used to compare the quality of models.^{124, 125} The AUC represents the probability that a randomly selected positive instance will be ranked more positive than a randomly selected negative one.¹²⁵

2.3.3.4 *Akaike information criterion (AIC)*

Another method for selection of a model among others is the Akaike information criterion. This criterion measures the relative quality of the fitted models.¹²⁶ Specifically, it helps the decision of how many descriptors to include in a model by introducing a penalty for adding variables to the model. This helps avoiding the problem of overfitting in the model. It is calculated using Equation (2-5):

$$AIC = 2k - 2\ln(L) \quad (2-5)$$

where k is the number of variables in the model and L is the maximized value of the likelihood function of the fitted model.¹²⁷

2.3.4 Graphical illustrations

Some figures in this work are presented in non-conventional plots. An example of these plots and what they mean is going to be shown in this section.

2.3.4.1 Receiver operating characteristic (ROC) curve

This curve was introduced by the British Royal Airforce back in World War II as a means of signal detection. Currently, it is being used in multiple fields of science.¹²⁸ In this thesis, we use the ROC curve for visualizing the performance of a binary classifier (logistic regression). This curve plots the false positive rate (fall-out) on the x-axis against the true positive rate (sensitivity) on y-axis at threshold levels between 0 and 1¹²⁹ as illustrated in Figure 2-10.

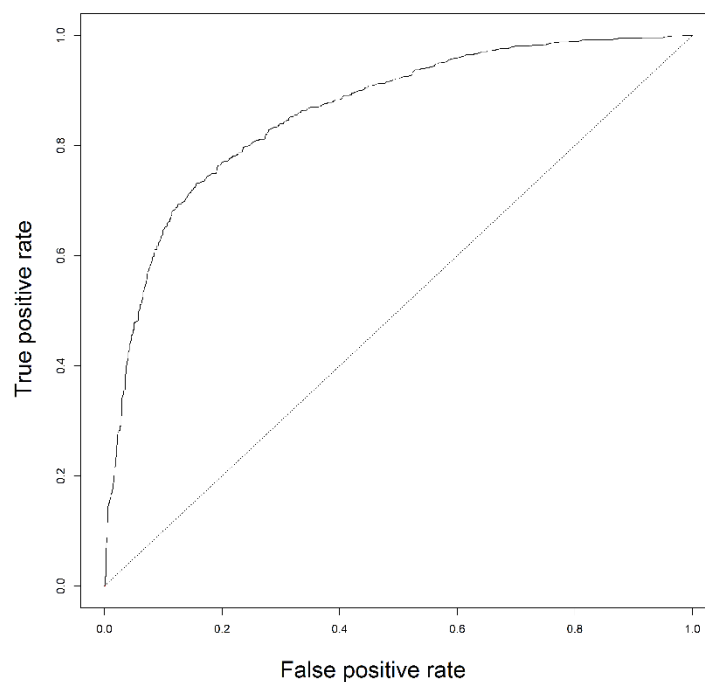


Figure 2-10. An example of the ROC curve of the chloroform model presented in this thesis.

Note that the diagonal line shown in the Figure 2-10 represents the random guess. Any model showing results above this line indicates a better performance. In the same graph it is possible to overlay the true positive rate (sensitivity) with the false positive rate (specificity). As seen in Figure 2-11.

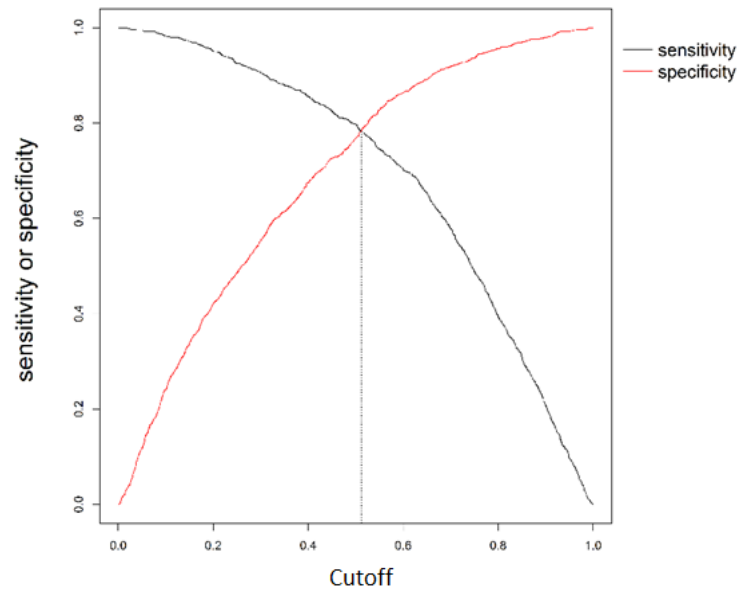


Figure 2-11. Sensitivity and specificity curves, with the optimal threshold level shown in the dotted line.

The threshold level that maximises both the sensitivity and specificity is the ideal point at which the model gives a non-biased prediction towards the positive or negative outcome.¹²⁸

2.3.4.2 *Boxplots*

Boxplots, also known as box and whiskers plot was introduced by Tukey.¹³⁰ This type of illustration visualizes variables values in terms of their quartiles. An example of this boxplot is shown in Figure 2-12.

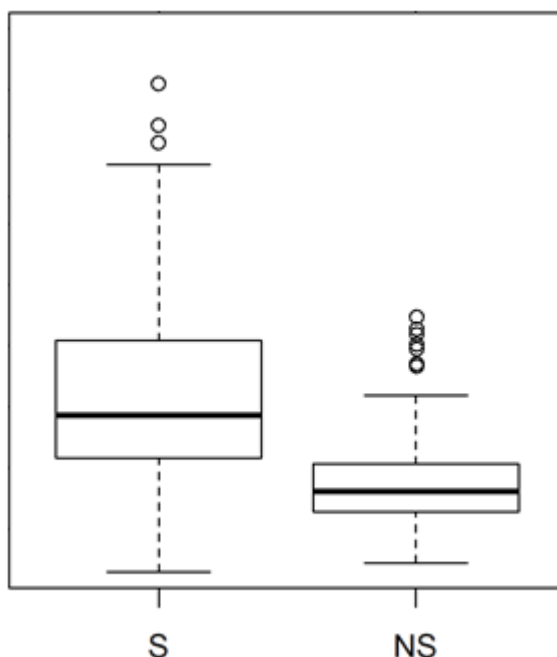


Figure 2-12. An example of a boxplot, obtained as a comparison between the solvate and the non-solvate groups in this thesis.

The lower side of the box represents the lower quartile. The middle of the box represents the median. The higher side of the box represents the upper quartile. The whiskers that extend outside the box represents a maximum of 1.5 of the interquartile range or a minimum of the most extreme point, whichever is closer to the median.¹³¹ outliers can be easily seen in this type of figure as they just sit outside the whiskers range.

2.4 Non-covalent interactions

In order to describe the structures that are included in this thesis, an insight into a few types of intermolecular interactions needs will be provided in sections 2.4.1 to 2.4.3.

2.4.1 Hydrogen bond

Hydrogen bonding is an attractive interaction between an electron-depleted hydrogen atom and an electron rich site, which can be represented by the scheme of $X-H\cdots Y$, where X,Y are

two atoms with greater electronegativity than the hydrogen atom H.¹³² The X-H part of the bond is known as the proton donor while the Y is known as the proton acceptor. The proton donor (also known as “donor” alone) can interact with multiple acceptors. Examples of atoms with greater electron density than hydrogen include C, O, N, F, P, Cl, Br, I, in addition to regions of high electron density such as double and triple bonds.¹³²

The first recognition of hydrogen bonding was in 1912 by T.S.Moore and T.F.Winmill¹³³ as mentioned by L. Pauling’s book in 1960.¹³⁴ Currently, the literature regarding hydrogen bonding is vast.^{135,136} However in this section we will focus on the aspects of hydrogen bond that have significant impact on the work conducted in this thesis.

In 2011, the IUPAC has released a document listing criteria for a bond to qualify as a hydrogen bond.¹³⁷ These criteria include almost linear (180 °) X-H...Y angle, prolonged X-H distance (evident from shifts of absorption bands as well as formation of new bands in IR spectra), deshielding of H (involved in the X-H bond) in NMR spectra. Additionally, the Gibbs energy related to the formation of the bond should be detectable experimentally.

The strength of a hydrogen bond can range from a 0.2 to 40 kcal/mol, making it second to ionic interactions.¹³⁶ In this wide energy range, hydrogen bonding can be broken down into 3 types: weak (1-5 kcal/mol), intermediate (5-15 kcal/mol) and strong hydrogen bonding (15-40 kcal/mol). The strength of the hydrogen bond can be estimated from its geometrical information (distances between atoms and angles between bonds).¹³⁸ Several reports have correlated the strength of hydrogen bonds to bond length and angle values acquired from large amount of spectroscopic or crystallographic data.^{138,139,140} These studies have allowed establishing guidelines of hydrogen bond distances and angles. Indicative distances¹³⁸ of some of the most common medium strong hydrogen bonds are given in Table 2-1. These are the main hydrogen bonds that have also been identified in the structures described in this work.

Table 2-1. Distances of some common hydrogen bonds

Bond type	H...A bond length	Sum of Van der Waals radii ^a (H...A)
O–H...O	1.215 to 1.230 Å (linear, strong bonds) to 2-3 Å (multi-centre bonds)	2.72
N–H...O	1.58 to 2.59 Å.	2.72
N–H...N	1.75–2.33 Å	2.75
O–H...N	1.59–2.19 Å	2.75

^a According to Bondi, taken from Mercury manual

From the examples demonstrated in Table 2-1, it can be seen that each hydrogen bond distance for medium strong bonds is shorter than the sum of van der Waals radii by at least ca. 0.1 Å. This was therefore selected as a limit for detecting hydrogen bonding during hydrogen bond analysis. However, it should be noted that hydrogen bonds are predominantly electrostatic interactions,^{141,142} therefore their angles and distances can vary in a long range. IUPAC in their recommendations have pointed out that the distances of strong hydrogen bonds would be shorter than the sum of van der Waals radii.¹³⁷ The angle should preferably be above 110 °, however, weak and multi-centre bonds can have smaller angles.

Alongside the medium strong hydrogen bonds given in Table 2-1, weak bonds such as C–H...O can also contribute to the stability of a crystal structure.^{138,143} The C...O separation for these bonds can be up to 4 Å and the energy of C–H...O hydrogen bonds is usually below 2 kcal/mol.¹⁴³

Hydrogen bonds contribute to the stability of crystalline structures and therefore are highly significant in solid-state chemistry. For example, hydrogen bonding is considered with regard to polymorphism of pharmaceutical ingredients.²⁸ Hydrogen bonds are also important in biochemistry as they ensure binding of ligands to the proteins.¹⁴⁴ Additionally, hydrogen bonding can direct chemical reactions¹⁴⁵ and impact physical properties of a solid.

2.4.2 Halogen bond

A halogen bond is an attractive interaction that forms between a halogen atom (e.g. F, Cl, Br, I) and an electron-rich site. It can be represented with the scheme $R-X\cdots Y$, where the X a halogen atom, Y is the electron rich site and R is a group that is covalently attached to the halogen X. Similar to hydrogen bonding, the electron-deficient site is known as the halogen bond donor (the R-X in the scheme) and the electron rich site is known as the halogen bond acceptor (the symbol Y in the scheme). This can get confusing because the former is sometimes referred to as the electron density acceptor and the latter is referred to as electron density donor. Note that a halogen atom X can form more than one halogen bond at the same time.^{146,147}

Halogen bond has been known since the late 19th century, where it was described by F. Guthrie as sticky electrophilic sites.¹⁴⁸ Later on, this type of bonding started receiving more attention.¹⁴⁹ The energy of this bond can range from 1.4 to 10 kcal/mol, as shown in an extensive study by S. Kozuch *et al.*¹⁵⁰ Due to this significant bond energy, it plays an important role structurally and biologically.^{151,152} The length of this bond is typically shorter than the sum of the van der Waals radii of the two interacting atoms.¹⁵³ In terms of its properties, this bond shares some properties with hydrogen bonding. For example, the driving force of the interaction is electrostatic. It is also a directional bond, where the interaction occurs at an angle of almost 180 ° considering the $R-X\cdots Y$ scheme. In fact, the Y approaches the X along the axis of the R-X bond. The interaction is thought to form following the σ -hole model, in which static charges plays a significant role.¹⁵⁴ Therefore, the more electron-deficient the halogen bond donor X is, the stronger the halogen bond would be.^{146,152}

2.4.3 Interactions of aromatic rings

In their simplest form, this type of interaction takes place between two unfunctionalized benzene rings, adopting one of three common arrangements known as sandwich (parallel),

offset-parallel and T-shaped (face-to-edge) interactions.¹⁵⁴ An illustration of these interactions is shown in Figure 2-13.

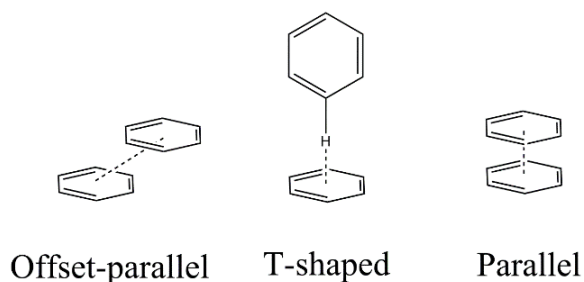


Figure 2-13. Graphic representation of ring interaction types.

The first observations of the (offset) parallel π - π interactions were made around 1930-1950 by Robertson and Lonsdale after analysing crystal structures of aromatic compounds.¹⁵⁵ Initially, π - π interactions were explained by formation of a charge transfer complex, but later experimental findings have shown that this explanation is unlikely.¹⁵⁵ The energy associated with π - π interactions is low and therefore difficult to determine. Several theoretical studies have been conducted to calculate the binding energy in benzene dimers, however, the method of calculation and the basis set can significantly affect the result. Sinnokrot *et al.*¹⁵⁴ compared the effect of different basis sets and concluded that aug-cc-pVDZ or larger basis set is necessary to obtain reliable results. The results obtained in their study are presented in Table 2-2. The authors have confirmed the correspondence of the geometry of benzene dimers obtained in their calculations to crystallographic observations.

Table 2-2. The distances and binding energy in the different π - π interactions

Interaction type	Distance between centroids, Å	Binding energy, kcal/mol
Sandwich	3.7	1.8
T-shaped	4.9	2.7
Parallel displaced	3.76 ^a	2.8

^a The reported distance is 3.6 vertically 1.6 horizontally (displacement), the value given is obtained by applying Pythagorean theorem.

The values reported in Table 2-2 shows the binding energies of the benzene dimer. Note that these conformations can have higher binding energies when the rings are substituted.¹⁵⁶ Ring interactions are known to contribute to the stability of biological systems, for example, the structure of DNA is supported by ring interactions.¹⁵⁷ These interactions also contribute to stability of crystal structures containing aromatic groups, such as pyrene.¹⁵⁸ Aromatic interactions also participate in binding ligands to proteins and therefore are a significant aspect of pharmaceutical compounds. For example, the complexation of the anaesthetic bupivacaine to receptors has been shown to rely on ring interactions.¹⁵⁹ In this work, we will focus on the role ring interactions play in solvate formation.

2.5 Characterization

2.5.1 X-ray diffraction techniques

In order understand solid-state materials it is crucial to obtain information about their structure. The most advantageous method in terms of information obtained is single crystal X-ray diffractometry. This method allows to obtain detailed information on the 3D structure of a material as well as to characterize its intermolecular interactions.^{160, 136} However, it is not always possible to use single-crystal X-ray diffractometry as it requires having single crystals of the material of appropriately large size. If use of single-crystal X-ray diffraction is not possible, crystal structure solution can be attempted from powder X-ray diffraction data. This method

gives less information where two dimensions of data are merged into one dimension. However, if sufficient information about the molecules in the crystal is available, it can be used for crystal structure determination. Powder X-ray diffraction is commonly used as a routine technique for identifying crystalline phases based on a “fingerprint” of peak positions.

Both diffraction methods are based on Bragg’s law which states that the distance d between Miller planes (planes in crystal structure formed by atoms or molecules) can be calculated from the 2θ position at which a reflection has been observed if the wavelength of radiation λ is known using Equation (2-6):

$$d = \frac{\lambda}{2\sin\theta} \quad (2-6)$$

2.5.1.1 Single crystal X-ray diffraction

In single crystal X-ray diffractometry a single crystal of the compound is analysed. X-ray diffraction patterns for this crystal are recorded depending on its angular position with regard to the source of radiation and detector. The intensity and position of diffraction peaks in these patterns contain information on atom positions and symmetry operations in the crystal lattice. The first step in extracting this information is indexing the diffraction pattern. Indexing assigns Miller indices to the diffraction peaks. This leads to characterizing the Miller planes orientation in the unit cell and allowing to calculate unit cell parameters.¹⁶¹ In further steps, the reflection intensities (corresponding to amplitude of the radiation wave) are used to calculate structure factor F_{hkl} , which contains information of the electron density in each point (x,y,z) of the unit cell. F_{hkl} can be calculated using Equation (2-7):

$$F_{hkl} = \int_V \rho(x, y, z) e^{2\pi i(hx + ky + lz)} dV \quad (2-7)$$

where h, k, l are Miller indices, x, y, z are fractional coordinates and V is volume of the unit cell. Fourier synthesis can be applied to calculate electron density,¹⁶² as presented in Equation (2-8):

$$\rho(x, y, z) = \frac{1}{V} \sum_h \sum_k \sum_l |F_{hkl}| e^{i\varphi_{hkl}} e^{-i2\pi(hx + ky + lz)} \quad (2-8)$$

This equation also contains a phase angle φ_{hkl} of the waves - this information is lost in the experiment causing the “phase problem”.^{162,163} Generally phase problem is solved by using the previous knowledge that crystals consist from discrete units – atoms – arranged in three-dimensional structures.

Several approaches have been developed to address the phase problem: direct methods,¹⁶³ charge flipping,¹⁶⁴ molecular replacement, VLD (Viva la difference)¹⁶⁵ etc. The most commonly used of these are direct methods, which employs structure invariants to link normalized structure factors to phase angles.¹⁶³ These algorithms are integrated in crystal structure determination and analysis packages such as SHELX, SIR, OLEX and other software packages.^{160,166}

2.5.1.2 Powder X-ray diffraction

Although crystal structure determination using powder X-ray diffraction data is considerably more complicated compared to single crystal X-ray diffraction, recent advances in computation techniques together with development of powerful algorithms have facilitated the use of this method.^{167,168} The first step in the crystal structure solution from powder data is generation of random trial structures, followed by simulation of their powder diffractograms and comparison of these with the experimental diffractogram. Subsequently, global optimum is found. The most commonly employed approach to find the global minimum is “Simulated Annealing” which is based on Monte Carlo method.¹⁶⁹

Other more common uses of X-ray powder diffraction include qualitative and quantitative analysis of solid phase mixtures. The distinctive property of X-ray powder diffraction is that it allows obtaining information about the structure of the material to be analysed, but not about its chemical composition therefore it is irreplaceable technique in polymorph screening.^{170,171} Nowadays, X-ray diffractometry is a fast method used in high-throughput solid form screening analysis. The data analysis is often conducted using statistical analysis – clustering and multivariate data analysis.¹⁷²

X-ray diffraction data can be used not only to identify solid phases present in a sample but also to quantitatively determine their amount. For quantitative analysis of X-ray diffraction data two approaches can be distinguished – individual reflection analysis and full pattern analysis. The first one is based on measuring the intensities of individual reflections with regard to a reference.¹⁷³ The full pattern analysis, on the other hand, is based on finding the best possible agreement between experimental and simulated X-ray diffraction patterns. The simulated pattern is calculated either from crystal structures (Rietveld method)¹⁷⁴ or from reference patterns of the components present in the sample.^{175,176,177}

2.5.2 Thermal analysis

Thermal methods such as thermogravimetry (TG), differential thermal analysis (DTA) and differential scanning calorimetry (DSC) are used to investigate the thermal behaviour of organic solids. Thermogravimetry is especially advantageous in identifying solvates, as this technique allows observing weight changes upon heating, which are often related to desolvation of solvates or decomposition of the material.¹⁷⁰ In order to analyse the by-products of decomposition, systems combining TG with evolved gas analysis have been developed. These systems commonly use mass spectrometry to analyse the products of decomposition.¹⁷⁸ DTA method is often used together with TG and allows identifying thermal events taking place upon heating the sample.¹⁷⁹

Thermogravimetry is extensively used in this work to identify solvate formation. A typical thermogravimeter consists from a programmable furnace and a balance.¹⁸⁰ Thermogravimetric measurement provides information on the temperature of the thermal event and the amount of lost weight.¹⁸¹ In addition, it is also often possible to differentiate between stoichiometric and nonstoichiometric solvates as the nature of the weight loss step is different for both of these. The stoichiometric solvates usually have well-defined weight loss steps in a narrow temperature range, while non-stoichiometric solvates lose weight in a wide temperature range and the weight loss may also vary for the same species.^{170, 182} A comparison of TG graphs comparing desolvation events of stoichiometric and non-stoichiometric solvates is shown in Figure 2-14.

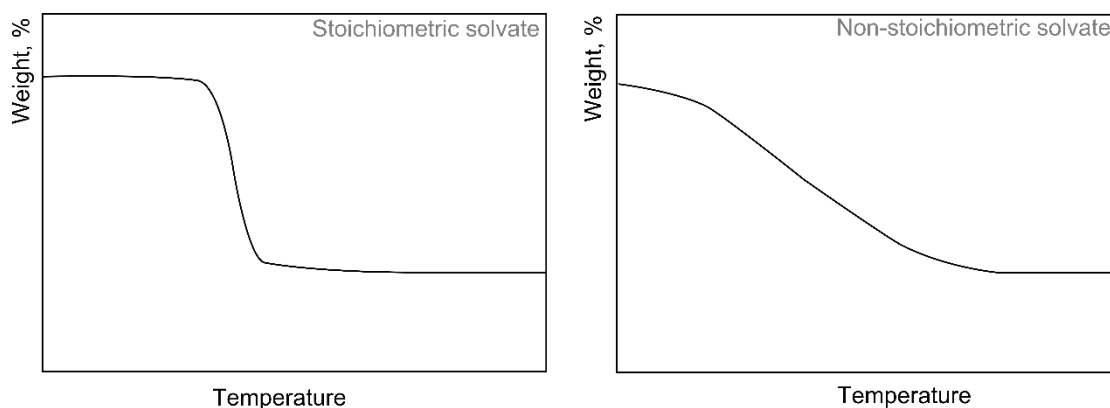


Figure 2-14. A comparison of desolvation events in TG thermograms for stoichiometric and non-stoichiometric solvates.

Although DSC is not going to be used in this work, it is worth highlighting the principles and uses of this method, as DSC and TG are considered two complementary methods. DSC detects the physical and chemical changes in a sample by measuring heat flow against a reference as a function of temperature and time. The output from DSC gives information about endo- and exothermic events, such as glass transition, crystallization and melting.^{183,184} Such properties make it especially useful in detection and characterization of a wide range of solid forms, including polymorphs, amorphous forms, multi-component solids as well as complete formulations. Due to its wide range of applications, it is reportedly one of the most heavily used methods in solids state characterization.¹⁸³ Temperature-modulated DSC (MTDSC), a relatively recent improvement to this method,¹⁸⁵ enables DSC to separate overlapping thermodynamic (irreversible) and kinetic (reversible) events.¹⁸⁶

2.6 References

1. Aitipamula S, Banerjee R, Bansal AK, Biradha K, Cheney ML, Choudhury AR, et al. Polymorphs, Salts, and Cocrystals: What's in a Name? *Crystal Growth & Design*. 2012;12(5):2147-52.
2. Grothe E, Meekes H, Vlieg E, ter Horst JH, de Gelder R. Solvates, salts and cocrystals: a proposal for a feasible classification system. *Crystal Growth & Design*. 2016;16(6):3237-43.
3. Vure P. Polymorph patents; how strong they are really? *International Journal of Intellectual Property Management*. 2011;4(4):297.
4. Petit S, Coquerel G. The Amorphous State. In: Hilfiker R, editor. *Polymorphism in the Pharmaceutical Industry*. Weinheim: Wiley-VCH; 2006. p. 259-84.
5. Babu NJ, Nangia A. Solubility Advantage of Amorphous Drugs and Pharmaceutical Cocrystals. *Crystal Growth & Design*. 2011;11(7):2662-79.
6. Kalepu S, Nekkanti V. Insoluble drug delivery strategies: Review of recent advances and business prospects. *Acta Pharmaceutica Sinica B*. 2015;5(5):442-53.
7. Li N, Ormes JD, Taylor LS. Leaching of Lopinavir Amorphous Solid Dispersions in Acidic Media. *Pharmaceutical Research*. 2016;33(7):1723-35.
8. Reddy BP, Reddy KR, Reddy DM, Reddy KS, Krishna BV, inventors. Amorphous form of lopinavir and ritonavir mixture. United States patent application US 14/003,535. 2012 Mar 5.
9. Jelińska A, Dudzińska I, Zając M, Oszczypowicz I. The stability of the amorphous form of cefuroxime axetil in solid state. *Journal of Pharmaceutical and Biomedical Analysis*. 2006;41(3):1075-81.
10. Crisp HA, Clayton JC, Elliott LG, Wilson EM, inventors; Glaxo Group Limited, assignee. Process for preparing cefuroxime axetil. United States patent US 5,013,833. 1991 May 7.

11. Rangineni S, Bhagwatwar H, Devarakonda S, Agarwal S, inventors; Bhagwatwar Harshal P, Devarakonda Surya N, Agarwal Sudeep K, assignee. Zafirlukast compositions. United States patent application US 11/671,480. 2007 Feb 6.
12. Yu L. Amorphous pharmaceutical solids: preparation, characterization and stabilization. *Advanced Drug Delivery Reviews*. 2001;48(1):27-42.
13. Ruland W. The Structure of Amorphous Solids. *Pure and Applied Chemistry*. 1969;18(4):489-516.
14. Willart JF, Descamps M. Solid state amorphization of pharmaceuticals. *Molecular Pharmaceutics*. 2008;5(6):905-20.
15. Einfalt T, Planinšek O, Hrovat K. Methods of amorphization and investigation of the amorphous state. *Acta Pharmaceutica (Zagreb, Croatia)*. 2013;63(3):305-34.
16. Halebian J, McCrone W. Pharmaceutical applications of polymorphism. *Journal of Pharmaceutical Sciences*. 1969;58(8):911-29.
17. Llinàs A, Goodman JM. Polymorph control: past, present and future. *Drug Discovery Today*. 2008;13(5-6):198-210.
18. Price SSL. Computed crystal energy landscapes for understanding and predicting organic crystal structures and polymorphism. *Accounts of Chemical Research*. 2009;42(1):117-26.
19. Lohani S, Grant DJW. Thermodynamic of Polymorphs. In: Hilfiker R, editor. *Polymorphism: In the pharmaceutical industry*. Weinheim: Wiley-VCH; 2006. p. 21-42.
20. Burger A, Ramberger R. On the Polymorphism of Pharmaceuticals and Other Molecular Crystals . II. *Mikrochimica Acta*. 1979;72(3):273-316.

21. Zhang GGZ, Zhou D. Crystalline and Amorphous Solids. In: Developing Solid Oral Dosage Forms: Pharmaceutical Theory and Practice. Elsevier; 2009. p. 25-60.
22. Grunenberg A, Henck JO, Siesler HW. Theoretical derivation and practical application of energy/temperature diagrams as an instrument in preformulation studies of polymorphic drug substances. *International Journal of Pharmaceutics*. 1996;129(1-2):147-58.
23. Burger A, Ramberger R. On the Polymorphism of Pharmaceuticals and Other Molecular Crystals. I. *Mikrochimica Acta*. 1979;72(3):259-71.
24. US Food Drug Administration. Regulatory Classification of Pharmaceutical Co-Crystals. Guidance for Industry. Center for Drug Evaluation and Research (CDER), Silver Spring, US. 2016.
25. Vishweshwar P, McMahon JA, Bis JA, Zaworotko MJ. Pharmaceutical Co-Crystals. *Journal of Pharmaceutical Sciences*. 2006;95(3):499-516.
26. Aakeröy CB, Salmon DJ. Building co-crystals with molecular sense and supramolecular sensibility. *CrystEngComm*. 2005;7(72):439-48.
27. Aakeröy CB, Fasulo ME, Desper J. Cocrystal or salt: does it really matter? *Molecular Pharmaceutics*. 2007;4(3):317-22.
28. Galek PTA, Fábián L, Allen FH. Persistent Hydrogen Bonding in Polymorphic Crystal Structures. *Acta Crystallographica Section B, Structural Science*. 2009;65:68-85.
29. Desiraju GR. A Bond by Any Other Name. *Angewandte Chemie International edition* . 2011;50(1):52 - 9.
30. Musumeci D, Hunter CA, Prohens R, Scuderi S, McCabe JF. Virtual cocrystal screening. *Chemical Science*. 2011;2(5):883-90.
31. Desiraju GR. Supramolecular Synthons in Crystal Engineering—A New Organic Synthesis. *Angewandte Chemie International Edition* . 1995;34(21):2311-27.

32. Allen FH. The Cambridge Structural Database: a quarter of a million crystal structures and rising. *Acta Crystallographica Section B, Structural Science*. 2002;B58:380-8.
33. Shattock TR, Arora KK, Vishweshwar P, Zaworotko MJ. Hierarchy of Supramolecular Synthons: Persistent Carboxylic Acid...Pyridine Hydrogen Bonds in Cocrystals That also Contain a Hydroxyl Moiety. *Crystal Growth & Design*. 2008;8(12):4533-45.
34. Allen FH, Samuel Motherwell WD, Raithby PR, Shields GP, Taylor R, . Systematic Analysis of the Probabilities of Formation of Bimolecular Hydrogen-Bonded Ring Motifs in Organic Crystal Structures. *New Journal of Chemistry*. 1999;23(1):25-34.
35. Griesser UJ. The Importance of Solvates. In: Hilfiker R, editor. *Polymorphism in the Pharmaceutical Industry*. Weinheim: Wiley-VCH; 2006. p. 211-33.
36. Vippagunta SR, Brittain HG, Grant DJW. Crystalline solids. *Advanced Drug Delivery Reviews*. 2001;48(1):3-26.
37. Morris KR. Structural aspects of hydrates and solvates. *Drugs and the Pharmaceutical Sciences*. 1999;95:125-81.
38. Price CP, Glick GD, Matzger AJ. Dissecting the Behavior of a Promiscuous Solvate Former. *Angewandte Chemie International Edition* . 2006;45(13):2062-6.
39. Tian F, Qu H, Zimmermann A, Munk T, Jørgensen AC, Rantanen J. Factors affecting crystallization of hydrates. *Journal of Pharmacy and Pharmacology*. 2010;62(11):1534-46.
40. van de Streek J, Rantanen J, Bond AD. Structures of cefradine dihydrate and cefaclor dihydrate from DFT-D calculations. *Acta Crystallographica Section C Crystal Structure Communications*. 2013;69:1229-33.
41. Zimmermann A, Tian F, de Diego HL, Frydenvang K, Rantanen J, Elema MR, et al. Structural Characterisation and Dehydration Behaviour of Siramesine Hydrochloride. *Journal of Pharmaceutical Sciences*. 2009;98(10):3596-607.

42. Stephenson GA, Groleau EG, Kleemann RL, Xu W, Rigsbee DR. Formation of isomorphic desolvates: Creating a molecular vacuum. *Journal of Pharmaceutical Sciences*. 1998;87(5):536-542.
43. Lee J, Boerrigter SXM, Jung YW, Byun Y, Yuk SH, Byrn SR, et al. Organic vapor sorption method of isostructural solvates and polymorph of tenofovir disoproxil fumarate. *European journal of pharmaceutical sciences* . 2013;50(3-4):253-562.
44. Bingham AL, Hughes DS, Hursthouse MB, Lancaster RW, Tavener S, Threlfall TL. Over one hundred solvates of sulfathiazole. *Chemical Communucations*. 2001;(7):603-604.
45. Clarke HD, Arora KK, Bass H, Kavuru P, Ong TT, Pujari T, et al. Structure–Stability Relationships in Cocrystal Hydrates: Does the Promiscuity of Water Make Crystalline Hydrates the Nemesis of Crystal Engineering? *Crystal Growth & Design*. 2010;10(5):2152-2167.
46. Khankari RK, Grant DJW. *Pharmaceutical Hydrates*. *Thermochimica Acta*. 1995;248:61-79.
47. Groenendaal JW, Leenderts EJ, Van Der Does T, inventors; Dsm Ip Assets BV, assignee. Crystalline amoxicillin trihydrate powder. United States patent application US 12/385,444. 2009 Apr 8.
48. Vermeersch HW, Thone DJ, Janssens LD, Wigerinck PT, inventors; Janssen R&D Ireland, assignee. Pseudopolymorphic Forms Of A HIV Protease Inhibitor. United States patent application US 14/183,712. 2014 Feb 19.
49. Gore VG, Patkar L, Bagul A, Vijaykar PS, Edake M, inventors; Generics [Uk] Limited assignee. Process for preparing crystalline dasatinib monohydrate. World patent application WO2010139980 A1, 2010.

50. Vacca JP, Lin JH, Yeh KC, Deutsch PJ, Ju WD, Chodakewitz JA, inventors; Merck & Co., Inc., assignee. Combination therapy for the treatment of AIDS. United States patent US 6,180,634. 2001 Jan 30.
51. Ghose AK, Herbertz T, Salvino JM, Mallamo JP. Knowledge-based chemoinformatic approaches to drug discovery. *Drug Discovery Today*. 2006;11(23–24):1107-14.
52. Manallack DT, Pitt WR, Gancia E, Montana JG, Livingstone DJ, Ford MG, et al. Selecting Screening Candidates for Kinase and G Protein-Coupled Receptor Targets Using Neural Networks. *Journal of Chemical Information and Computer Sciences*. 2002;42(5):1256-62.
53. Morissette SL, Almarsson O, Peterson ML, Remenar JF, Read MJ, Lemmo AV, et al. High-throughput crystallization: polymorphs, salts, co-crystals and solvates of pharmaceutical solids. *Advanced Drug Delivery Reviews*. 2004;56(3):275-300.
54. Pereira DA, Williams JA. Origin and evolution of high throughput screening. *British Journal of Pharmacology*. 2007;152(1):53-61.
55. Szymański P, Markowicz M, Mikiciuk-Olasik E. Adaptation of High-Throughput Screening in Drug Discovery—Toxicological Screening Tests. *International Journal of Molecular Sciences*. 2012;13(1):427-52.
56. Wölcke J, Ullmann D. Miniaturized HTS technologies - uHTS. *Drug Discovery Today*. 2001;6(12):637-46.
57. Reilly AM, Cooper RI, Adjiman CS, Bhattacharya S, Boese AD, Brandenburg JG, et al. Report on the sixth blind test of organic crystal structure prediction methods. *Acta Crystallographica Section B*. 2016;72(4):439-59.
58. Day GM, Cooper TG, Cruz-Cabeza AJ, Hejczyk KE, Ammon HL, Boerrigter SXM, et al. Significant progress in predicting the crystal structures of small organic molecules - a report on the fourth blind test. *Acta Crystallographica Section B, Structural Science*. 2009;65(2):107-25.

59. Price SL. Predicting crystal structures of organic compounds. *Chemical Society Reviews*. 2014;43(7):2098-111.
60. Price SL, Braun DE, Reutzel-Edens SM. Can computed crystal energy landscapes help understand pharmaceutical solids? *Chemical Communications*. 2016;52(44):7065-77.
61. Price SL. Computational prediction of organic crystal structures and polymorphism. *International Reviews in Physical Chemistry*. 2008;27(3):541-68.
62. Lommerse JPM, Motherwell WDS, Ammon HL, Dunitz JD, Gavezzotti A, Hofmann DWM, et al. A test of crystal structure prediction of small organic molecules. *Acta Crystallographica Section B, Structural Science*. 2000;56(4):697-714.
63. Motherwell WDS, Ammon HL, Dunitz JD, Dzyabchenko A, Erk P, Gavezzotti A, et al. Crystal structure prediction of small organic molecules: a second blind test. *Acta Crystallographica Section B, Structural Science*. 2002;58(4):647-61.
64. Day GM, Motherwell WDS, Ammon HL, Boerrigter SXM, Della Valle RG, Venuti E, et al. A third blind test of crystal structure prediction. *Acta Crystallographica Section B, Structural Science*. 2005;61(5):511-27.
65. Bardwell DA, Adjiman CS, Arnautova YA, Bartashevich E, Boerrigter SXM, Braun DE, et al. Towards crystal structure prediction of complex organic compounds – a report on the fifth blind test. *Acta Crystallographica Section B, Structural Science*. 2011;67(6):535-51.
66. Willett P. *Chemoinformatics: a history*. Wiley Interdisciplinary Reviews: Computational Molecular Science. 2011;1(1):46-56.
67. Brown FK. Chapter 35 - Chemoinformatics: What is it and How does it Impact Drug Discovery. In: James AB, editor. *Annual Reports in Medicinal Chemistry*. Volume 33: Academic Press; 1998. p. 375-84.

68. Wold S. Chemometrics; what do we mean with it, and what do we want from it? Chemometrics and Intelligent Laboratory Systems. 1995;30(1):109-15.
69. Luscombe NM, Greenbaum D, Gerstein M. What is Bioinformatics? A Proposed Definition and Overview of the Field. Methods of Information in Medicine. 2001;40(4):346-58.
70. Nantasenamat C, Isarankura-Na-Ayudhya C, Naenna T, Prachayasittikul V. A PRACTICAL OVERVIEW OF QUANTITATIVE STRUCTURE-ACTIVITY RELATIONSHIP EXCLI Journal. 2009;8:74-88.name is written in capital letters in the original paper
71. Brown AC. On the theory of Chemical Combination: University of Edinburgh; 1861.
72. Brown AC, Fraser TR. On the Connection between Chemical Constitution and Physiological Action; with special reference to the Physiological Action of the Salts of the Ammonium Bases derived from Strychnia, Brucia, Thebaia, Codeia, Morphia, and Nicotia. Journal of Anatomy and Physiology. 1868;2(2):224-42.
73. Duffy BC, Zhu L, Decornez H, Kitchen DB. Early phase drug discovery: Cheminformatics and computational techniques in identifying lead series. Bioorganic & Medicinal Chemistry. 2012;20(18):5324-42.
74. Todeschini R, Consonni V. Handbook of Molecular Descriptors: Wiley-VCH; 2000.
75. Talete srl, Dragon (Software for Molecular Descriptor Calculation) <http://www.talete.mi.it/>. 6.0 ed2014.
76. Bingham AL, Hughes DS, Hursthouse MB, Lancaster RW, Tavener S, Threlfall TL. Over one hundred solvates of sulfathiazole. Chem Commun [Internet]. 2001;(7):603–4.
77. Price CP, Glick GD, Matzger AJ. Dissecting the behavior of a promiscuous solvate former. Angew Chemie - Int Ed. 2006;45(13):2062–6.

78. Sarceviča I, Grante I, Belyakov S, Rekis T, Bērziņš K, Actiņš A, et al. Solvates of Dasatinib: Diversity and Isostructurality. *J Pharm Sci* [Internet]. 2016 Apr;105(4):1489–95.
79. Infantes L, Fábíán L, Motherwell WDS. Organic crystal hydrates: what are the important factors for formation. *CrystEngComm*. 2007;9(1):65.
80. Nangia A, Desiraju GR. Pseudopolymorphism : occurrences of hydrogen bonding organic solvents in molecular crystals. 1999;(February):605–6.
81. Gillon AL, Feeder N, Davey RJ, Storey R. Hydration in molecular crystals - A Cambridge Structural Database analysis. *Cryst Growth Des*. 2003;3(5):663–73.
82. Brychczynska M, Davey RJ, Pidcock E. A study of methanol solvates using the Cambridge structural database. *New J Chem* [Internet]. 2008;32(10):1754–60.
83. Van de Streek J, Motherwell S. New software for searching the Cambridge Structural Database for solvated and unsolvated crystal structures applied to hydrates. *CrystEngComm*. 2007;9(1):55.
84. Van de Streek J, All series of multiple solvates (including hydrates) from the Cambridge Structural Database. *CrystEngComm* [Internet]. 2007;9(5):350–2.
85. Whitley E, Ball J. Statistics review 3: Hypothesis testing and P values. *Critical Care*. 2002;6(3):222-5.
86. Mann HB, Whitney DR. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*. 1947;18(1):50-60.
87. Lowry R. Concepts and applications of inferential statistics. 2003.
88. Motulsky HJ. Prism 5 Statistics Guide. San Diego CA: GraphPad Software Inc.; 2007.
89. Wilcoxon F. Individual Comparisons by Ranking Methods. *Biometrics Bulletin*. 1945;1(6):80-3.

90. Lehmann EL, Romano JP. Testing statistical hypotheses: SpringerTexts in Statistics; 2006.
91. Hubbard R. Alphabet Soup: Blurring the Distinctions Betweenp's anda's in Psychological Research. *Theory & Psychology*. 2004;14(3):295-327.
92. Dorey F. In Brief: The P Value: What Is It and What Does It Tell You? *Clinical Orthopaedics and Related Research®*. 2010;468(8):2297-8.
93. Dahiru T. P-Value, a true test of statistical significance? a cautionary note. *Annals of Ibadan postgraduate medicine*. 2008;6(1):21-6.
94. Hand D, Mannila H, Smyth P. Principles of Data Mining: The MIT Press; 2001.
95. Cyril G, Cancedda N, Dymetman M, Foster G. Learning Machine Translation. MIT Press, Cambridge, USA; 2009.
96. Michie D, Spiegelhalter DJ, Taylor CC. Machine Learning, Neural and Statistical Classification. 1994.
97. Abdi H, Williams LJ. Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*. 2010;2(4):433-59.
98. Pearson K. LIII. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine Series 6*. 1901;2(11):559-72.
99. Jackson JE. PCA With More Than Two Variables. In: A user's guide to principal components. 1991:26-62.
100. Dray S. On the number of principal components: A test of dimensionality based on measurements of similarity between matrices. *Computational Statistics & Data Analysis*. 2008;52(4):2228-37.

101. Josse J, Husson F. Selecting the number of components in principal component analysis using cross-validation approximations. *Computational Statistics & Data Analysis*. 2012;56(6):1869-79.
102. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Vienna, Austria: R Foundation for Statistical Computing; 2015.
103. Kotsiantis SB, Zaharakis ID, Pintelas PE. Machine learning: a review of classification and combining techniques. *Artificial Intelligence Review*. 2006;26(3):159-90.
104. Hanrahan G. Artificial neural networks in biological and environmental analysis 1st Edition ed: Boca Raton, FL : CRC Press. ; 2011.
105. David G. Kleinbaum MK. Link. Logistic Regression: A Self-Learning Tex; 2010.
106. Cortes C, Vapnik V. Support-vector networks. *Machine Learning*. 1995;20(3):273-97.
107. Cox DR. The Regression Analysis of Binary Sequences. *Journal of the Royal Statistical Society Series B (Methodological)*. 1958;20(2):215-42.
108. Peng C-YJ, Lee KL, Ingersoll GM. An Introduction to Logistic Regression Analysis and Reporting. *The Journal of Educational Research*. 2002;96(1):3-14.
109. Bewick V, Cheek L, Ball J. Statistics review 14: Logistic regression. *Critical Care*. 2005;9(1):112-8.
110. Mountrakis G, Im J, Ogole C. Support vector machines in remote sensing: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*. 2011;66(3):247-59.
111. Osuna E, Freund R, Girosi F. Support vector machines: Training and applications. 1997.
112. Chen Y, Councill IG. An Introduction to Support Vector Machines: A Review. *AI MAGAZINE*. 2003;24(2):105-6.

113. Burbidge R, Buxton B. An Introduction to Support Vector Machines for Data Mining. 2001:3-15.
114. Gaspar P, Carbonell J, Oliveira JL. On the parameter optimization of Support Vector Machines for binary classification. *Journal of Integrative Bioinformatics*. 2012;9(3):201.
115. Hsu C-W, Chang C-C, Lin C-J. A Practical Guide to Support Vector Classification. 2003.
116. Veropoulos K, Campbell C, Cristianini N. Controlling the sensitivity of support vector machines. In *Proceedings of the international joint conference on AI 1999 Jul 31* (pp. 55-60).
117. Krogh A, Vedelsby J, editors. *Neural Network Ensembles, Cross Validation, and Active Learning*. *Advances in Neural Information Processing Systems 7*; MTM Press; 1995:231-8.
118. Arlot S, Celisse A. A survey of cross-validation procedures for model selection. 2010:40-79.
119. Mosier CI. The need and means of cross validation. I. Problems and designs of cross-validation. *Educational and Psychological Measurement*. 1951;11:5-11.
120. Wold S. Cross-Validatory Estimation of the Number of Components in Factor and Principal Components Models. *Technometrics*. 1978;20(4):397-405.
121. Browne MW. Cross-Validation Methods. *Journal of Mathematical Psychology*. 2000;44(1):108-32.
122. Kohavi R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2*; Montreal, Quebec, Canada. 1643047: Morgan Kaufmann Publishers Inc.; 1995. p. 1137-43.
123. Eldar YC. Uniformly Improving the Cramér-Rao Bound and Maximum-Likelihood Estimation. *IEEE Transactions on Signal Processing*. 2006;54(8):2943-56.

124. Bamber D. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology*. 1975;12(4):387-415.
125. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 1982;143(1):29-36.
126. Akaike H. Factor analysis and AIC. *Psychometrika*. 1987;52(3):317-32.
127. Akaike H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*. 1974;19(6):716-23.
128. Carter JV, Pan J, Rai SN, Galandiuk S. ROC-ing along: Evaluation and interpretation of receiver operating characteristic curves. *Surgery*. 2016;159(6):1638-45.
129. Lusted LB. Signal Detectability and Medical Decision-Making. *Science*. 1971;171(3977):1217-9.
130. Tukey JW. *Exploratory Data Analysis: Past, Present and Future*. Defence Technical Information Center Document, 1993.
131. McGill R, Tukey JW, Larsen WA. Variations of Box Plots. *The American Statistician*. 1978;32(1):12-6.
132. Kollman PA, Allen LC. Theory of the hydrogen bond. *Chemical Reviews*. 1972;72(3):283-303.
133. Moore TS, Winmill TF. CLXXVII.—The state of amines in aqueous solution. *Journal of the Chemical Society, Transactions*. 1912;101:1635-76.
134. Pauling L. *The nature of the chemical bond and the structure of molecules and crystals: an introduction to modern structural chemistry*: Cornell university press; 1960.

135. Bernstein J, Etter MC, Leiserowitz L. The Role of Hydrogen Bonding in Molecular Assemblies. *Structure Correlation*: Wiley-VCH Verlag GmbH; 2008. p. 431-507.
136. Steiner T. The Hydrogen Bond in the Solid State. *Angewandte Chemie International Edition*. 2002;41(1):48-76.
137. Arunan E, Desiraju GR, Klein RA, Sadlej J, Scheiner S, Alkorta I, et al. Definition of the hydrogen bond (IUPAC Recommendations 2011). *Pure and Applied Chemistry*. 2011;83(8):1637-41.
138. Jeffrey GA. Hydrogen-bonding: An update. *Crystallography Reviews*. 2003;9(2-3):135-76.
139. Novak A. Hydrogen bonding in solids correlation of spectroscopic and crystallographic data. *Large Molecules*: Springer; 1974;(18): 177-216.
140. Taylor R, Kennard O. Hydrogen-bond geometry in organic crystals. *Accounts of Chemical Research*. 1984; 17(9):320–6.
141. Hunter CA. Quantifying Intermolecular Interactions: Guidelines for the Molecular Recognition Toolbox. *Angewandte Chemie International Edition*. 2004;43(40):5310-24.
142. Gavezzotti A. The Crystal Packing of Organic Molecules: Challenge and Fascination Below 1000 Da. 1998. p. 5-121.
143. Desiraju GR. The C–H···O hydrogen bond in crystals: what is it? *Accounts of Chemical Research*. 1991; 24(10):290-296.
144. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*. 1983;22(12):2577-637.

145. Hine J, Linden SM, Kanagasabapathy V. Double-hydrogen-bonding catalysis of the reaction of phenyl glycidyl ether with diethylamine by 1, 8-biphenylenediol. *The Journal of Organic Chemistry*. 1985;50(25):5096-9.
146. Desiraju GR, Ho PS, Kloo L, Legon AC, Marquardt R, Metrangolo P, et al. Definition of the halogen bond (IUPAC Recommendations 2013). *Pure and Applied Chemistry*. 2013;85(8):1711-3.
147. Metrangolo P, Resnati G. Halogen Versus Hydrogen. *Science*. 2008;321(5891):918-9.
148. Guthrie F. XXVIII.—On the iodide of iodammonium. *Journal of the Chemical Society*. 1863;16(0):239-44.
149. Mulliken RS. Structures of Complexes Formed by Halogen Molecules with Aromatic and with Oxygenated Solvents1. *Journal of the American Chemical Society*. 1950;72(1):600-8.
150. Kozuch S, Martin JML. Halogen Bonds: Benchmarks and Theoretical Analysis. *Journal of Chemical Theory and Computation*. 2013;9(4):1918-31.
151. Hassel O. Structural Aspects of Interatomic Charge-Transfer Bonding. *Science*. 1970;170(3957):497-502.
152. Scholfield MR, Zanden CMV, Carter M, Ho PS. Halogen bonding (X-bonding): A biological perspective. *Protein Science*. 2013;22(2):139-52.
153. Rowland RS, Taylor R. Intermolecular Nonbonded Contact Distances in Organic Crystal Structures: Comparison with Distances Expected from van der Waals Radii. *The Journal of Physical Chemistry*. 1996;100(18):7384-91.
154. Sinnokrot MO, Valeev EF, Sherrill CD. Estimates of the Ab Initio Limit for π - π Interactions: The Benzene Dimer. *Journal of the American Chemical Society*. 2002;124(36):10887-93.

155. Dahl T. The Nature of Stacking Interactions Between Organic Molecules Elucidated by Analysis of Crystal Structures. *Acta Chemica Scandinavica*. 1994;48(2):95-106.
156. Smith T, Slipchenko LV, Gordon MS. Modeling π - π Interactions with the Effective Fragment Potential Method: The Benzene Dimer and Substituents. *The Journal of Physical Chemistry A*. 2008;112(23):5286-94.
157. Mignon P, Loverix S, Steyaert J, Geerlings P. Influence of the π - π interaction on the hydrogen bonding capacity of stacked DNA/RNA bases. *Nucleic Acids Research*. 2005;33(6):1779-89.
158. Robertson JM, White JG. 72. The crystal structure of pyrene. A quantitative X-ray investigation. *Journal of the Chemical Society (Resumed)*. 1947(0):358-68.
159. Powell E, Lee YH, Partch R, Dennis D, Morey T, Varshney M. Pi-Pi complexation of bupivacaine and analogues with aromatic receptors: implications for overdose remediation. *International Journal of Nanomedicine*. 2007;2(3):449-59.
160. Sheldrick GM. A short history of SHELX. *Acta Crystallographica Section A, Foundations of Crystallography*. 2008;64:112-22.
161. Kabsch W. Automatic Indexing of Rotation Diffraction Patterns. *Journal of Applied Crystallography*. 1988;21:67-72.
162. Taylor G. The phase problem. *Acta Crystallographica - Section D Biological Crystallography*. 2003;59(Pt 11):1881-90.
163. Hauptman H. The phase problem of x-ray crystallography. *Proceedings of the Indian Academy of Sciences - Chemical Sciences*. 1983; 92 (4-5): 291-321.
164. Oszlányi G, Süto A. The charge flipping algorithm. *Acta Crystallographica Section A, Foundations of Crystallography*. 2008;64:123-34.

165. Burla MC, Giacovazzo C, Polidori G. From a Random to the Correct Structure: the VLD Algorithm. *Journal of Applied Crystallography*. 2010;43:825-36.
166. Sheldrick GM. SHELXT – Integrated space-group and crystal-structure determination. *Acta Crystallographica Section A Foundations and Advances*. 2015;71:3-8.
167. David WIF, Shankland K. Structure Determination from Powder Diffraction Data. *Acta Crystallographica Section A Foundations of Crystallography*. 2008;64:52-64.
168. Karki S, Fábíán L, Frišćić T, Jones W. Powder X-ray Diffraction as an Emerging Method to Structurally Characterize Organic Solids. *Organic Letters*. 2007;9(16):3133-6.
169. Zhukov SG, Chernyshev VV, Babaev EV, Sonneveld EJ, Schenk H. Application of simulated annealing approach for structure solution of molecular crystals from X-ray laboratory powder data. *Zeitschrift fuer Kristallographie - Crystalline Materials*. 2001;216(1):5-9.
170. Brittain HG. Methods for the characterization of polymorphs and solvates. *Drugs and the Pharmaceutical Sciences*. 1999;95:227-78.
171. Newman A. X-ray powder diffraction in solid form screening and selection. *American Pharmaceutical Review*. 2011;14:44-51.
172. Barr G, Dong W, Gilmore CJ. High-throughput powder diffraction . II . Applications of clustering methods and multivariate data analysis . *Journal of Applied Crystallography*. 2004;37:243-52.
173. Padrela L, de Azevedo EG, Velaga SP. Powder X-ray Diffraction Method for the Quantification of Cocrystals in the Crystallization Mixture. *Drug Development and Industrial Pharmacy*. 2012;38(8):923-9.
174. Bish DL, Howard SA. Quantitative Phase Analysis Using the Rietveld Method. *Journal of Applied Crystallography*. 1988;21:86-91.

175. Gilmore CJ, Barr G, Paisley J. High-Throughput Powder Diffraction. I. A New Approach to Qualitative and Quantitative Powder Diffraction Pattern Analysis Using Full Pattern Profiles. *Journal of Applied Crystallography*. 2004;37:231-42.
176. Chipera SJ, Bish DL. FULLPAT : a Full-Pattern Quantitative Analysis Program for X-Ray Powder Diffraction Using Measured and Calculated Patterns. *Journal of Applied Crystallography*. 2002;35:744-9.
177. Chipera SJ, Bish DL. Fitting Full X-Ray Diffraction Patterns for Quantitative Analysis : A Method for Readily Quantifying Crystalline and Disordered Phases. *Advances in Materials Physics and Chemistry*. 2013;3(1):47-53.
178. Kamruddin M, Ajikumar PK, Dash S, Tyagi AK, Raj B. Thermogravimetry-evolved gas analysis-mass spectrometry system for materials research. *Bulletin of Materials Science*. 2003;26(4):449-60.
179. Barrall EM, Gernert JF, Porter RS, Johnson JF. Differential Thermal Analysis Apparatus. *Analytical Chemistry*. 1963;35:1837-40.
180. Coats AW, Redfern JP. Thermogravimetric analysis. A review. *Analyst*. 1963;88:906-24.
181. Earnest CM. Modern Thermogravimetry. *Analytical Chemistry*. 1984;56:1471A-86A.
182. Craig DQM, Galwey AK. Thermogravimetric Analysis. In: Craig DQM, Reading M, editors. *Thermal Analysis of Pharmaceuticals*: CRC Press; 2006. p. 139-92.
183. Chieng N, Rades T, Aaltonen J. An overview of recent studies on the analysis of pharmaceutical polymorphs. *Journal of Pharmaceutical and Biomedical Analysis*. 2011;55(4):618-44.
184. Gill P, Moghadam TT, Ranjbar B. Differential Scanning Calorimetry Techniques: Applications in Biology and Nanoscience. *Journal of Biomolecular Techniques*. 2010;21(4):167-93.

185. Reading á. Modulated differential scanning calorimetry—a new way forward in materials characterization. *Trends Polym Sci.* 1993;1(8):248-53.

186. Kawakami K, Ida Y. Application of modulated-temperature DSC to the analysis of enantiotropically related polymorphic transitions. *Thermochimica Acta.* 2005;427(1):93-9.

Chapter 3: **Materials and methods**

3.1 Materials

The solvents used in this work were ethanol ($\text{C}_2\text{H}_6\text{O}$, Sigma Aldrich, UK, purity: $\geq 99\%$ (GC)), methanol (CH_3O , Fisher Scientific, UK, purity: HPLC grade), dichloromethane (CH_2Cl_2 , Sigma Aldrich, UK, purity: $\geq 99\%$ (GC)), chloroform (CHCl_3 , Fisher Scientific, UK, purity: laboratory reagent grade) as well as Milli-Q water that is prepared on site in the School of Pharmacy at the University of East Anglia.

The pharmaceutically active ingredients that were used were Theophylline anhydrous ($\text{C}_7\text{H}_8\text{N}_4\text{O}_2$, Sigma Aldrich, UK), 4-Methylumbelliferone (Hymecromone) ($\text{C}_{10}\text{H}_8\text{O}_3$, Sigma Aldrich, UK), Isoniazid (isonicotinohydrazide) ($\text{C}_6\text{H}_7\text{N}_3\text{O}$, Fluka, $\geq 99\%$, India), Ethenzamide (2-ethoxybenzamide) ($\text{C}_9\text{H}_{11}\text{NO}_2$, Alfa Aesar, Germany), Carbamazepine ($\text{C}_{15}\text{H}_{12}\text{N}_2\text{O}$, Sigma Aldrich, UK), Diflunisal ($\text{C}_{13}\text{H}_8\text{F}_2\text{O}_3$, Sigma Aldrich UK), Fenofibrate ($\text{C}_{20}\text{H}_{21}\text{ClO}_4$, generously donated by Merck Serono (Germany)), Felodipine ($\text{C}_{18}\text{H}_{19}\text{NO}_4\text{Cl}_2$, Afine Chemicals Limited, Hangzhou, China), Ketoconazole ($\text{C}_{26}\text{H}_{28}\text{Cl}_2\text{N}_4\text{O}_4$, Alfa Aesar, Germany), Griseofulvin ($\text{C}_{17}\text{H}_{17}\text{ClO}_6$, Alfa Aesar, Germany).

3.2 Methods

3.2.1 Descriptor calculation

Descriptor calculation was conducted using the Dragon software. Relevant technical details of calculation process are given below.

3.2.1.1 Main descriptors calculated by Dragon

The Dragon software that was used in this thesis calculates 29 families of descriptors, known as blocks. These blocks are presented in Table 3-1.

Table 3-1. The 29 main blocks of descriptors calculated by Dragon. Details explaining the calculation of each descriptor can be found in the Dragon software documentation¹

Block number	Block name	Examples (Name in Dragon)
1	Constitutional descriptors	Molecular weight (MW), number of non-H atoms (nSK), percentage of O atoms (O %)
2	Ring descriptors	Number of rings (nCIC), number of 6-membered rings (nR06), total ring size (TRS)
3	Topological indices	First Zagreb index (ZM1) ² , all path Wiener index (Wap) ³

Table 3-1. Continued

4	Walk and path counts	Molecular path count of order 1 (MPC01, nr. of bonds between non-H atoms); self-returning walk count of order 4 (SRW04, nr. of possible 4-bond walks that return to the same atom)
5	Connectivity indices	Randic connectivity index (X1) ⁴ , solvation connectivity index of order 0 (X0sol) ⁵
6	Information indices	The information index on molecular size (ISIZ) ⁶ , Kier symmetry index (S0K) ⁴
7	2D matrix-based descriptors	Average vertex sum from reciprocal squared distance matrix (AVS_H2), third order spectral moment of the topological distance matrix (SM3_D)
8	2D autocorrelations	Moran autocorrelation of lag 1 weighted by mass (MATS1m) ⁷
9	Burden matrix eigenvalues ⁸	Largest eigenvalue of Burden matrix weighted by mass (SpMax1_Bhm)
10	P_VSA-like descriptors ⁹	P_VSA-like descriptor on LogP, 1st bin (P_VSA_LogP_1)

Table 3-1. Continued

11	ETA indices ¹⁰	Eta core count (Eta_alpha), eta p shape index (Eta_sh_p)
12	Edge adjacency indices	Leading eigenvalue from the edge adjacency matrix of the H-depleted molecular graph (SpMax_EA)
13	Geometrical descriptors	radius of gyration (Rgyr), gravitational index G1 (G1) ¹¹
14	3D matrix-based descriptors	Wiener-like index from distance/distance matrix (Wi_G/D) ^{12 13}
15	3D autocorrelations	3D Topological distance based autocorrelation with lag 1, unweighted (TDB01u) ¹⁴
16	RDF descriptors	Value of the Radial Distribution Function weighted by atomic mass at 1.5 Å (RDF015m) ¹⁵
17	3D-MoRSE descriptors ¹⁶	Unweighted scattered intensity at a scattering ratio of 1 (Mor01u)
18	WHIM descriptors ¹⁷	Unweighted variance of atomic coordinates along the 1st principal axis of the molecule (L1u)
19	GETAWAY descriptors ^{18, 19}	Atomic leverage weighted autocorrelation of lag 0 (H0u)
20	Randic molecular profiles ²⁰	Molecular profile 1 (sum of all interatomic distances, divided by the number of atoms, DP01)

Table 3-1. Continued

21	Functional group counts	Number of aliphatic carboxyl groups (nRCOOH)
22	Atom-centred fragments ²¹	Number of CH ₂ R ₂ fragments in the molecule (C-002)
23	Atom-type E-state indices ²²	Sum of the electrotopological states of all methyl carbon atoms in the molecule (SsCH ₃)
24	CATS 2D ^{23, 24}	Number of donor-donor atom pairs at a separation of 3 bond (CATS2D_03_DD)
25	2D Atom Pairs ²⁵	Sum of topological distances between N..N atom pairs (T(N..N))
26	3D Atom Pairs	Sum of geometrical distances between N..N atom pairs (G(N..N))
27	Charge descriptors	The maximum positive atomic charge (qpmax)
28	Molecular properties	The hydrophilic factor (Hy) ²⁶ , total surface area of acceptor atoms (SAacc)
29	Drug-like indices	Complementary Lipinski Alert index (cRo5, 0 or 1 indicating whether the molecule violates at least two criteria from Lipinski's rule of five) ^{27, 28}

3.2.1.2 The reciprocal squared topological distance matrix (H_2)

This matrix is mentioned when the ethanol solvate formation model is presented. The process of calculating topological distance matrix (H_2) will be described using a commonly used drug; aspirin (2-(acetoxy)benzoic acid). Calculations for other compounds are conducted similarly.

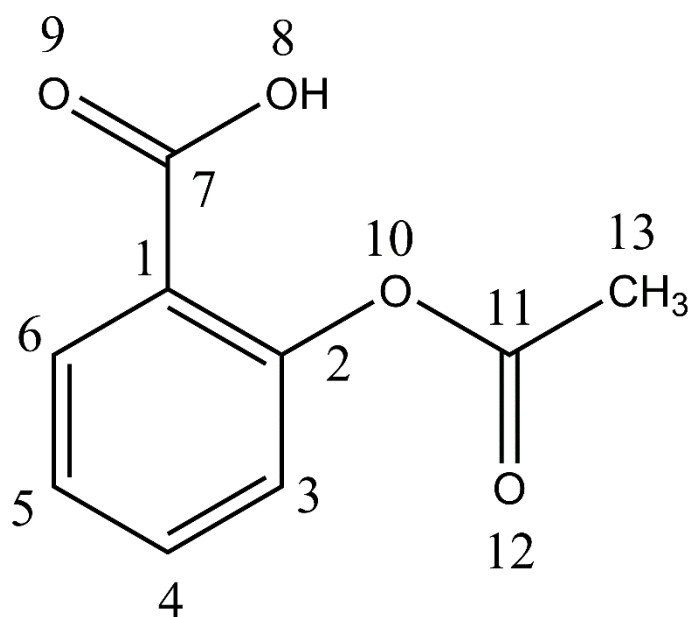


Figure 3-1. Molecular graph of the aspirin molecule.

The topological distance matrix is calculated from the molecular graph. The topological distance between atoms i and j (d_{ij}) is defined as the number of bonds in the shortest path connecting them in the molecular graph. The diagonal elements of the topological distance matrix are zero, while the off-diagonals give the topological distance between non-hydrogen atoms. An example for aspirin is shown in Table 3-2 (refer to Figure 3-1 for atom numbering).

Table 3-2. Topological distance matrix of aspirin

-	1	2	3	4	5	6	7	8	9	10	11	12	13
1	0	1	2	3	2	1	1	2	2	2	3	4	4
2	1	0	1	2	3	2	2	3	3	1	2	3	3
3	2	1	0	1	2	3	3	4	4	2	3	4	4
4	3	2	1	0	1	2	4	5	5	3	4	5	5
5	2	3	2	1	0	1	3	4	4	4	5	6	6
6	1	2	3	2	1	0	2	3	3	3	4	5	5
7	1	2	3	4	3	2	0	1	1	3	4	5	5
8	2	3	4	5	4	3	1	0	2	4	5	6	6
9	2	3	4	5	4	3	1	2	0	4	5	6	6
10	2	1	2	3	4	3	3	4	4	0	1	2	2
11	3	2	3	4	5	4	4	5	5	1	0	1	1
12	4	3	4	5	6	5	5	6	6	2	1	0	2
13	4	3	4	5	6	5	5	6	6	2	1	2	0

Squaring each element in the matrix gives the squared topological distance matrix, shown in

Table 3-3.

Table 3-3. Squared topological distance matrix of aspirin

-	1	4	9	16	25	36	49	64	81	100	121	144	169
1	0	1	4	9	4	1	1	4	4	4	9	16	16
4	1	0	1	4	9	4	4	9	9	1	4	9	9
9	4	1	0	1	4	9	9	16	16	4	9	16	16
16	9	4	1	0	1	4	16	25	25	9	16	25	25
25	4	9	4	1	0	1	9	16	16	16	25	36	36
36	1	4	9	4	1	0	4	9	9	9	16	25	25
49	1	4	9	16	9	4	0	1	1	9	16	25	25
64	4	9	16	25	16	9	1	0	4	16	25	36	36
81	4	9	16	25	16	9	1	4	0	16	25	36	36
100	4	1	4	9	16	9	9	16	16	0	1	4	4
121	9	4	9	16	25	16	16	25	25	1	0	1	1
144	16	9	16	25	36	25	25	36	36	4	1	0	4
169	16	9	16	25	36	25	25	36	36	4	1	4	0

By taking the reciprocal of each element in the squared matrix, the reciprocal squared topological distance matrix (H2) is obtained as shown in Table 3-4.

Table 3-4. Reciprocal squared topological distance matrix (H2) of aspirin

0.00	1.00	0.25	0.11	0.06	0.04	0.03	0.02	0.02	0.01	0.01	0.01	0.01	0.01
1.00	0.00	1.00	0.25	0.11	0.25	1.00	1.00	0.25	0.25	0.25	0.11	0.06	0.06
0.25	1.00	0.00	1.00	0.25	0.11	0.25	0.25	0.11	0.11	1.00	0.25	0.11	0.11
0.11	0.25	1.00	0.00	1.00	0.25	0.11	0.11	0.06	0.06	0.25	0.11	0.06	0.06
0.06	0.11	0.25	1.00	0.00	1.00	0.25	0.06	0.04	0.04	0.11	0.06	0.04	0.04
0.04	0.25	0.11	0.25	1.00	0.00	1.00	0.11	0.06	0.06	0.06	0.04	0.03	0.03
0.03	1.00	0.25	0.11	0.25	1.00	0.00	0.25	0.11	0.11	0.11	0.06	0.04	0.04
0.02	1.00	0.25	0.11	0.06	0.11	0.25	0.00	1.00	1.00	0.11	0.06	0.04	0.04
0.02	0.25	0.11	0.06	0.04	0.06	0.11	1.00	0.00	0.25	0.06	0.04	0.03	0.03
0.01	0.25	0.11	0.06	0.04	0.06	0.11	1.00	0.25	0.00	0.06	0.04	0.03	0.03
0.01	0.25	1.00	0.25	0.11	0.06	0.11	0.11	0.06	0.06	0.00	1.00	0.25	0.25
0.01	0.11	0.25	0.11	0.06	0.04	0.06	0.06	0.04	0.04	1.00	0.00	1.00	1.00
0.01	0.06	0.11	0.06	0.04	0.03	0.04	0.04	0.03	0.03	0.25	1.00	0.00	0.25
0.01	0.06	0.11	0.06	0.04	0.03	0.04	0.04	0.03	0.03	0.25	1.00	0.25	0.00

3.2.1.3 Main descriptors used in thesis

The values of these descriptors for two common drugs are illustrated in Table 3-5. The descriptors that require previous knowledge of the 3D structure of a molecule were not included.

Table 3-5. Values of the descriptors mentioned in the thesis calculated for carbamazepine and acetaminophen molecules

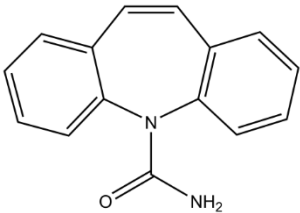
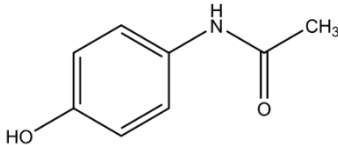
Descriptor	Brief explanation of the descriptor		
AVS_H2	The log transformation of the average of the sum of the entries in each row of the reciprocal squared topological distance matrix.	3.783	3.016
TRS	The number of atoms in each independent ring in the molecule.	19	6
SM3_H2	The log transformation of the third order spectral moment of the reciprocal squared distance matrix.	4.168	3.358
π ID	The logarithmic transform of the conventional bond order ID number.	9.776	6.033

Table 3-5. Continued

nHDon	The number of hydrogen bond donors.	2	2
Hy	The hydrophilic factor	0.32	0.66
H-050	The number of hydrogen atoms attached to a heteroatom	2	2
nH	The number of hydrogen atoms.	12	9
nCIC	The number of rings in the molecule.	3	1
MPC01	The log transformation of the count of paths of length 1 in the H-depleted molecular graph.	3.045	2.485

3.2.2 Slurry preparation

Chapter 7 of this thesis incorporated screening of materials for solvate and hydrate formation. All slurries were prepared using one general procedure, that is, the addition of an amount of the solid material into a 1-1.5 ml of the solvent until slurry is formed. Five slurries were prepared per candidate, each of them being with a different solvent (ethanol, methanol, water, chloroform, dichloromethane), resulting in a total of 50 slurries. After these slurries were prepared, they were left to shake at 25 °C and 250 rpm for 20 days. The details of each sample preparation are shown in Table 3-6.

Table 3-6. A summary of the preparation of each slurry used in the screening experiments

Drug candidate	Solvent	Volume of solvent added (ml)	Amount of drug added (mg)	Drying time (minutes)
Theophylline	Ethanol	1.5	50	20
Theophylline	Methanol	1.5	50	20
Theophylline	Dichloromethane	1	50	10
Theophylline	Chloroform	1	50	10
Theophylline	Water	1.5	50	60
Hymecromone	Ethanol	1.5	50	20
Hymecromone	Methanol	1.5	50	20
Hymecromone	Dichloromethane	1	50	10
Hymecromone	Chloroform	1	50	10
Hymecromone	Water	1.5	50	60
Griseofulvin	Ethanol	1.5	50	20
Griseofulvin	Methanol	1.5	50	20
Griseofulvin	Dichloromethane	1	250	10
Griseofulvin	Chloroform	1	100	10
Griseofulvin	Water	1.5	50	60
Isoniazid	Ethanol	1.5	100	20
Isoniazid	Methanol	1.5	100	20
Isoniazid	Dichloromethane	1	100	10
Isoniazid	Chloroform	1	100	10
Isoniazid	Water	1.5	200	60
Ethenzamide	Ethanol	1.5	100	20
Ethenzamide	Methanol	1.5	150	20
Ethenzamide	Dichloromethane	1	200	10

Table 3-6. Continued

Ethenzamide	Chloroform	1	200	10
Ethenzamide	Water	1.5	50	60
Carbamazepine	Ethanol	1.5	100	20
Carbamazepine	Methanol	1.5	150	20
Carbamazepine	Dichloromethane	1	250	10
Carbamazepine	Chloroform	1	250	10
Carbamazepine	Water	1.5	50	60
Diflunisal	Ethanol	1.5	300	20
Diflunisal	Methanol	1.5	250	20
Diflunisal	Dichloromethane	1	50	10
Diflunisal	Chloroform	1	50	10
Diflunisal	Water	1.5	100	60
Fenofibrate	Ethanol	1.5	100	20
Fenofibrate	Methanol	1.5	100	20
Fenofibrate	Dichloromethane	1	750	10
Fenofibrate	Chloroform	1	750	10
Fenofibrate	Water	1.5	50	60
Felodipine	Ethanol	1.5	150	20
Felodipine	Methanol	1.5	180	20
Felodipine	Dichloromethane	0.5	650	10
Felodipine	Chloroform	0.5	600	10
Felodipine	Water	1.5	50	60
Ketoconazole	Ethanol	1.5	80	20
Ketoconazole	Methanol	1.5	150	20
Ketoconazole	Dichloromethane	1	500	10
Ketoconazole	Chloroform	1	500	10

Table 3-6. Continued

Ketoconazole	Water	1.5	50	60
--------------	-------	-----	----	----

3.2.3 Thermogravimetric analysis

In this work, the thermogravimetric analysis was conducted to detect the solvate formation. TGA Q5000 (TA instruments, Newcastle, USA) was used for this purpose. The samples were prepared from slurry then dried at room temperature. After drying they were heated from 40 °C up to 250 °C at a rate of 10 °C min⁻¹. The nitrogen gas purge rate was set to 100 ml min⁻¹.

3.2.4 Single crystal X-ray diffraction.

The single crystal X-ray diffraction experiments in this thesis were performed using an Oxford Diffraction Xcalibur-3/Sapphire3-CCD diffractometer (Oxford diffraction Ltd., Oxford, UK) equipped with a graphite monochromator. The diffractometer uses a Mo-K α radiation of wavelength 0.71073 Å. Intensity data was measured by thin-slice ω - and ϕ -scans. All experiments were conducted at 140(1) K. The programs CrysAlisPro-CCD and -RED were used to process the diffraction data.^{29,30} The structures were solved in SHELXT³¹ *via* the dual-space approach. SHELXL and the user interface ShelXle were used to refine the structures.^{32,33} All non-hydrogen atoms were located from electron density maps. The thermal displacement parameters of these atoms were refined anisotropically. Hydrogen atoms were added in geometrically idealized positions and their coordinates were refined in riding mode, while allowing rigid rotations of the methyl groups. All hydrogen atoms were refined with isotropic displacement parameters.

Crystal structure analysis and geometric measurements were carried out using the programs PLATON and OLEX2.^{34, 35} The graphical illustrations were created using Mercury.³⁶

3.2.5 PXRD

Powder X-ray Diffraction experiments were conducted using a Thermo-ARL X'tra diffractometer (Ecublens, Switzerland). Cu K α 1 radiation source was used with 45 kV voltage and a current of 40 mA. All experiments were conducted at room temperature and humidity.

PXRD patterns of hymecromone, diflunisal and fenofibrate were obtained in this work. All samples were prepared by gentle crushing of the crystals using a metal spatula or a mortar and pestle and transferring the crushed material into the sample holder. The diffraction data was recorded in 2 θ -range from 3 ° to 50 °. The measurement was carried out at a rate of one second per step and the step size of 0.01 °. Note that the solvate samples were measured immediately after crushing, to avoid possible solvent loss.

3.2.6 Microscopy

The hot-stage microscopy experiments in reflective mode were carried out using LinkamMDSG600 automated hot stage and Linkam imaging station that was attached to a microscope with LED light source and $\times 10$ magnification lens.

The polarised light microscopy experiments were conducted using Leica DM LS2 polarised light microscope (Wetzlar GmbH, Germany) connected to a video capture system and equipped with the Mettler Toledo FP 82 HT hot stage and FP 90 temperature controller.

3.3 References

1. Talete srl, Dragon (Software for Molecular Descriptor Calculation). <http://www.taletе.mi.it/>. 6.0 ed2014.
2. Gutman I, Rušćić B, Trinajstić N, Wilcox CF. Graph theory and molecular orbitals. XII. Acyclic polyenes. The Journal of Chemical Physics. 1975;62(9):3399-405.
3. Lukovits I. An All-Path Version of the Wiener Index. Journal of Chemical Information and Computer Sciences. 1998;38(2):125-9.
4. Thomson G. Molecular connectivity in structure-activity analysis. By L. B. Kier and L. H. Hall, Research studies Press (a division of John Wiley and Sons), Letchworth, Herefordshire, England. AIChE Journal. 1987;33(12):2096.
5. Zefirov NS, Palyulin VA. QSAR for Boiling Points of "Small" Sulfides. Are the "High-Quality Structure-Property-Activity Regressions" the Real High Quality QSAR Models? Journal of Chemical Information and Computer Sciences. 2001;41(4):1022-7.
6. Bertz SH. The first general index of molecular complexity. Journal of the American Chemical Society. 1981;103(12):3599-601.
7. Moran PA. Notes on continuous stochastic phenomena. Biometrika. 1950;37(1-2):17-23.
8. Pearlman RS, Smith KM. Novel Software Tools for Chemical Diversity. In: Kubinyi H, Folkers G, Martin Y, editors. 3D QSAR in Drug Design. Three-Dimensional Quantitative Structure Activity Relationships. 2: Springer Netherlands; 1998. p. 339-53.

9. Labute P. A widely applicable set of descriptors. *Journal of Molecular Graphics and Modelling*. 2000;18(4–5):464-77.
10. Roy K, Ghosh G. Introduction of Extended Topochemical Atom (ETA) Indices in the Valence Electron Mobile (VEM) Environment as Tools for QSAR/QSPR Studies. *Internet Electronic Journal of Molecular Design*. 2003;2:599-620.
11. Katritzky AR, Mu L, Lobanov VS, Karelson M. Correlation of Boiling Points with Molecular Structure. 1. A Training Set of 298 Diverse Organics and a Test Set of 9 Simple Inorganics. *The Journal of Physical Chemistry*. 1996;100(24):10400-7.
12. Randic M, Kleiner AF, De Alba LM. Distance/Distance Matrixes. *Journal of Chemical Information and Computer Sciences*. 1994;34(2):277-86.
13. Hosoya H. Topological Index. A Newly Proposed Quantity Characterizing the Topological Nature of Structural Isomers of Saturated Hydrocarbons. *Bulletin of the Chemical Society of Japan*. 1971;44(9):2332-9.
14. Klein CT, Kaiser D, Ecker G. Topological Distance Based 3D Descriptors for Use in QSAR and Diversity Analysis. *Journal of Chemical Information and Computer Sciences*. 2004;44(1):200-9.
15. Hemmer MC, Steinhauer V, Gasteiger J. Deriving the 3D structure of organic molecules from their infrared spectra. *Vibrational Spectroscopy*. 1999;19(1):151-64.
16. Gasteiger J, Sadowski J, Schuur J, Selzer P, Steinhauer L, Steinhauer V. Chemical Information in 3D Space. *Journal of Chemical Information and Computer Sciences*. 1996;36(5):1030-7.

17. Todeschini R, Lasagni M, Marengo E. New molecular descriptors for 2D and 3D structures. Theory. Journal of Chemometrics. 1994;8(4):263-72.
18. Consonni V, Todeschini R, Pavan M. Structure/Response Correlations and Similarity/Diversity Analysis by GETAWAY Descriptors. 1. Theory of the Novel 3D Molecular Descriptors. Journal of Chemical Information and Computer Sciences. 2002;42(3):682-92.
19. Consonni V, Todeschini R, Pavan M, Gramatica P. Structure/Response Correlations and Similarity/Diversity Analysis by GETAWAY Descriptors. 2. Application of the Novel 3D Molecular Descriptors to QSAR/QSPR Studies. Journal of Chemical Information and Computer Sciences. 2002;42(3):693-705.
20. Randic M. Molecular Shape Profiles. Journal of Chemical Information and Computer Sciences. 1995;35(3):373-82.
21. Viswanadhan VN, Ghose AK, Revankar GR, Robins RK. Atomic physicochemical parameters for three dimensional structure directed quantitative structure-activity relationships. 4. Additional parameters for hydrophobic and dispersive interactions and their application for an automated superposition of certain naturally occurring nucleoside antibiotics. Journal of Chemical Information and Computer Sciences. 1989;29(3):163-72.
22. Hall LH, Kier LB, Brown BB. Molecular Similarity Based on Novel Atom-Type Electrotopological State Indices. Journal of Chemical Information and Computer Sciences. 1995;35(6):1074-80.
23. Fechner U, Franke L, Renner S, Schneider P, Schneider G. Comparison of correlation vector methods for ligand-based similarity searching. Journal of Computer-Aided Molecular Design. 2003;17(10):687-98.

24. Schneider G, Neidhart W, Giller T, Schmid G. "Scaffold-Hopping" by Topological Pharmacophore Search: A Contribution to Virtual Screening. *Angewandte Chemie International Edition*. 1999;38(19):2894-6.
25. Carhart RE, Smith DH, Venkataraghavan R. Atom pairs as molecular features in structure-activity studies: definition and applications. *Journal of Chemical Information and Computer Sciences*. 1985;25(2):64-73.
26. Todeschini R, Vighi M, Finizio A, Gramatica P. 3D-modelling and prediction by WHIM descriptors. Part 8. Toxicity and physico-chemical properties of environmental priority chemicals by 2D-TI and 3D-WHIM descriptors. *SAR and QSAR in Environmental Research*. 1997;7(1-4):173-93.
27. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews*. 1997;23(1-3):3-25.
28. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews*. 2001;46(1-3):3-26.
29. Agilent (2014). CrysAlis PRO. Agilent Technologies Ltd, Yarnton, Oxfordshire, England.
30. Oxford Diffraction (2006). CrysAlis CCD, CrysAlis PRO. Oxford Diffraction Ltd, Abingdon, Oxfordshire, England.
31. Sheldrick GM. SHELXT – Integrated space-group and crystal-structure determination. *Acta Crystallographica Section A Foundations and Advances*. 2015;71:3-8.

32. Sheldrick GM. Crystal structure refinement with SHELXL. *Acta Crystallographica Section C: Structural Chemistry*. 2015 Jan 1;71(1):3-8.
33. Hübschle CB, Sheldrick GM, Dittrich B. ShelXle: a Qt graphical user interface for SHELXL. *Journal of Applied Crystallography*. 2011;44(6):1281-4.
34. Spek AL. Structure validation in chemical crystallography. *Acta Crystallographica Section D: Biological Crystallography*. 2009;65(2):148-55.
35. Dolomanov OV, Bourhis LJ, Gildea RJ, Howard JAK, Puschmann H. OLEX2: a complete structure solution, refinement and analysis program. *Journal of Applied Crystallography*. 2009;42(2):339-41.
36. Macrae CF, Bruno IJ, Chisholm JA, Edgington PR, McCabe P, Pidcock E, Rodriguez-Monge L, Taylor R, Streek JV, Wood PA. Mercury CSD 2.0—new features for the visualization and investigation of crystal structures. *Journal of Applied Crystallography*. 2008 Apr 1;41(2):466-70.

Chapter 4: Data acquisition and descriptor calculation

4.1 Overview of the research steps

The introduction in Chapter 1 has given a background on how the work in this thesis is designed to predict the hydrate and solvate formation in organic compounds. An overview of the technical steps that need to be taken to achieve the desired predictions is summarized in Figure 4-1.

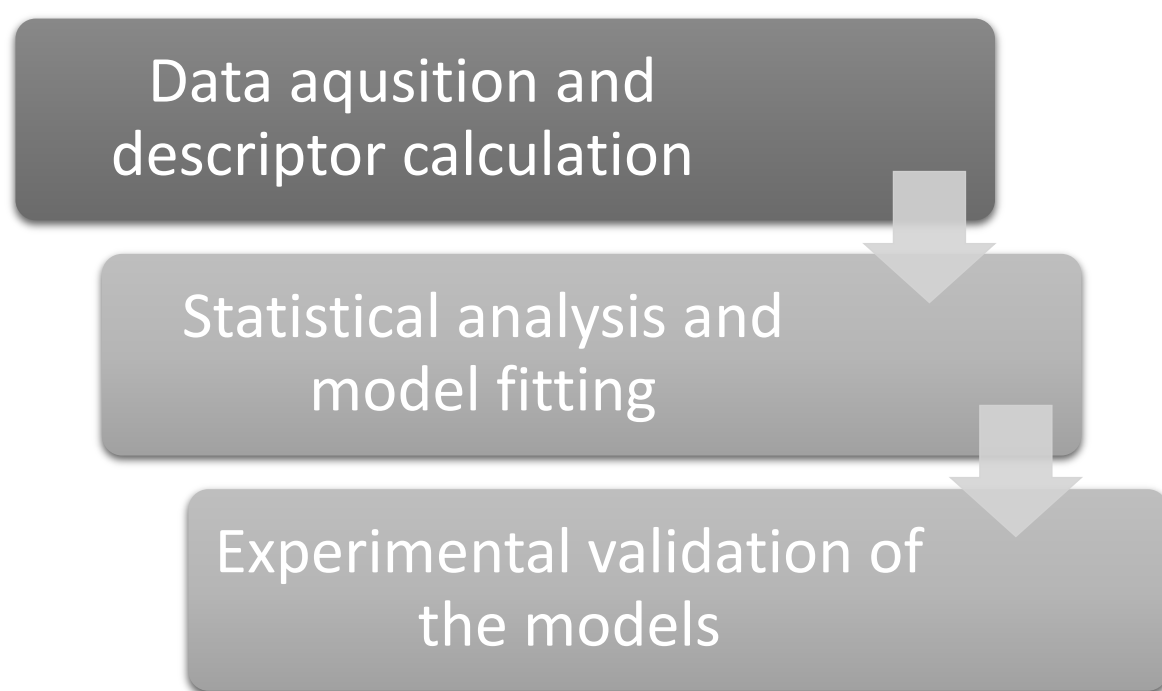


Figure 4-1. Main steps of analysis of solvate formation.

4.2 Data collection

In this section, technical details on how the data was collected, the challenges faced and how each issue was treated will be discussed.

4.2.1 Solvent selection rationale

The steps presented in Figure 4-1 seems to be in a logical order, but the question remains what solvents should be studied for their solvate formation ability? Over 300 recrystallization solvents are recorded in the CSD, as shown by a study conducted in 2000.¹ Studying this large

number would not be feasible due to the time limit. Additionally, it would not be possible to draw reliable conclusions for solvents having small number of hits in the database. For these reasons, few solvents had to be selected. Solvents with the largest number of hits in the database would be a rational choice, as this is a knowledge based approach, where the larger amount of information could lead to more robust results.

The same study that reported the 300+ solvents has also shown the top 50 solvents ranked by number of solvates they formed with organic compounds in the CSD.¹ Note that the word solvate here means solvate that is formed by a single solvent, for example a dichloromethane solvate. The information in this article was published based on the CSD version of October 1998. To check if these numbers are still valid at this time, the top 10 solvents addressed in the article were checked using the CSD November 2013 version. The comparison can be seen in Figure 4-2.

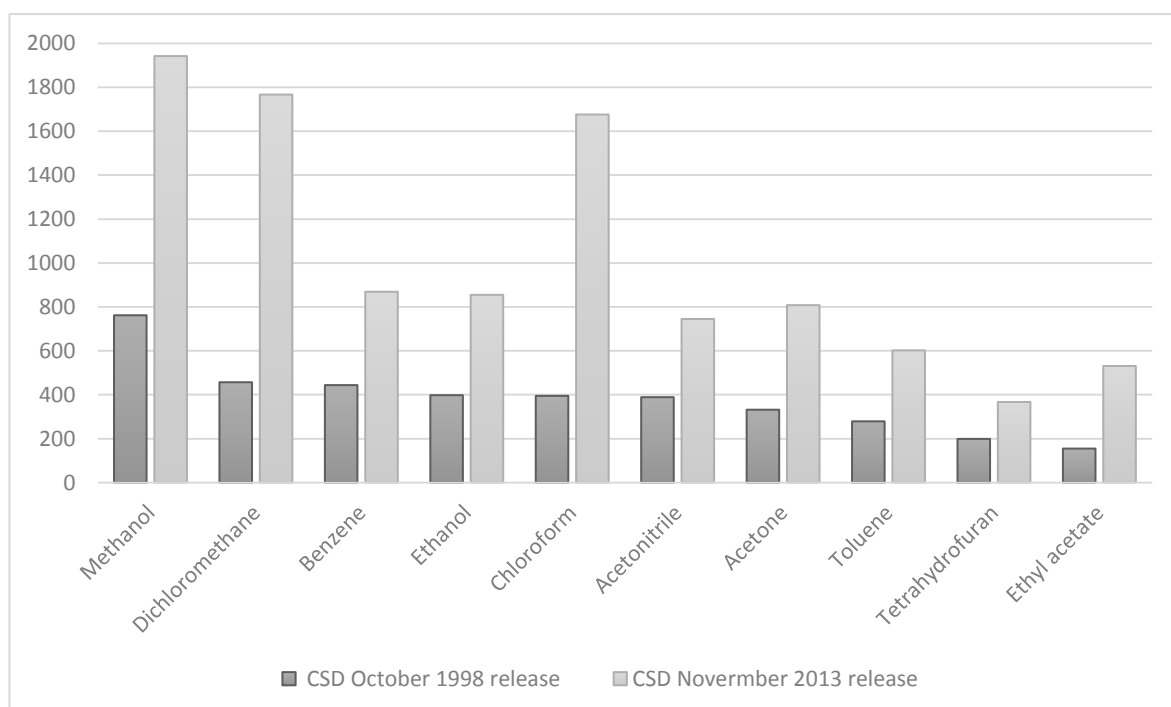


Figure 4-2. The number of solvates recorded in CSD for the 10 most commonly used organic recrystallization solvents.

The top 5 solvate-forming organic solvents (with organic materials) in 1998 remained as the top forming ones in 2013, these solvents are methanol, dichloromethane, benzene, ethanol and chloroform. With exception to benzene, all of these solvents are reported to be commonly used solvents in pharmaceutical industry.² Additionally, ethanol and methanol belong to the same family of hydroxyl group-containing solvents, similarly, dichloromethane and chloroform are also closely related chlorinated solvents. Comparing the hydroxylated to the chlorinated solvents could have several benefits on the study. For example, solvents of the same group are expected to have similar behaviour. Benzene did not show the same potential as it stands in a different group. For these reasons, ethanol, methanol, dichloromethane and chloroform were selected. Water, being the most commonly used solvent and the largest solvate former, where hydrates represent around 33 % of entries in the CSD was also included in this study.³ Although dichloromethane and chloroform have well recognized high toxicity^{4,5}, a large increase (around 4-folds) in the number of dichloromethane and chloroform solvates among organic crystals was observed, relative to the year 1998. This shows the increasing usage of these solvents in crystallization; therefore the importance of studying them.

4.2.2 Entries selection rationale

Identification of the factors that contribute to solvate formation in the five chosen solvents is possible through comparing the molecules which were able to form a solvate to those that were not able to form a solvate, both being crystallized from the same solvent. In order to do this comparison, two groups of molecules per solvent need to be extracted from the database, these are: a solvate forming and a non-solvate forming group. In order to keep the study in a defined shape, it was necessary to apply restrictions on the selected molecules. For example, formally charged molecules are expected to form an ionic bond, which is significantly stronger than any other non-covalent interaction.⁶ This shows that the inclusion of such entries would cause an increase in the number of outliers in the data. The ConQuest software was used to

make a custom search in the database. The solvate-forming entries for each solvent were selected based on the following criteria:

- Each entry should have 2 chemical entities that are not covalently bonded.
- One of these 2 entities is the solvent of interest.
- Only organic structures were considered, no organometallic structures were searched.
- Structures that are ionic (salts) or polymeric were also excluded from the search.

To extract the non-solvate entries that were recrystallized from a certain solvent, the search criteria were set to:

- Each entry should have 1 molecule only.
- The recrystallization solvent should only be the solvent of interest.
- Only organic structures were considered, no organometallic structures were searched.
- Structures that are ionic (salts) or polymeric were excluded from the search.

Note that in both groups, only organic structures were selected. This is due to the fact that the reason for solvent inclusion (esp. water) changes with the presence of metal ions. This is reported to occur due to the entry of water atoms into the coordination sphere of metal atoms.⁷

A total of 54,653 CSD entries were exported using the search criteria explained. The extracted data consisted of 15,082 solvate-forming (S) and 39,571 non-solvate-forming (NS) structures. The breakdown of these entries between the 5 solvents is shown in Table 4-1.

Table 4-1. The breakdown of the number of solvate and non-solvate entries by solvent

Solvent	NS	S	Total
Ethanol	17,958	855	18,813
Methanol	7,862	1,942	9,804
Dichloromethane	7,149	1,767	8,916
Chloroform	4,382	1,676	6,058
Water	2,220	8,842	11,062
Total	39,571	15,082	54,653

4.2.3 Refinement of the non-solvate/solvate forming groups

Random samples of the non-solvate and solvate groups were visually examined in each solvent to ensure they meet the desired criteria mentioned above. Five types of unexpected entries were encountered.

(1) When the non-solvate forming groups were searched for, ConQuest text search was used. It was intended that entries recrystallized from the solvent of interest only would be selected from the database. When the entries were investigated, it was seen that many of them had a combination of recrystallization solvents. Additionally, entries having the solvent as part of their name were also within the results of text search. An example of that is 2-(Boranyl(*t*-butyl)methylphosphoranyl)-1,1-diphenylethanol (CSD refcode: BETMAQ)⁸ which was listed as a result for the search, although it doesn't have ethanol as the recrystallization solvent. Moreover, an overlap in the names of the solvents was observed. For example, the string "ethanol" was used to search the recrystallization solvent of each entry in the database. Results for molecules that were recrystallized from methanol also turned up. This explains the

large number of entries under the ethanol non-solvate group compared to the other organic solvents' groups. It is notable that some of these issues could have been avoided if the recrystallization solvent search was used instead of the text search in the first place. However, these issues were fixed by processing the non-solvate lists through a Linux command line script that strictly looks for the recrystallization solvent of interest in each dataset.

(2) Upon examining the solvate-forming groups, it was noticed that some of the entries had no recrystallization solvent recorded. These entries were part of the results of the search because the search criteria for solvate-forming entries did not include a criterion to look in the recrystallization solvent tab in the database. The percentage of entries with unknown recrystallization solvent versus the complete solvate dataset is shown in Figure 4-3.

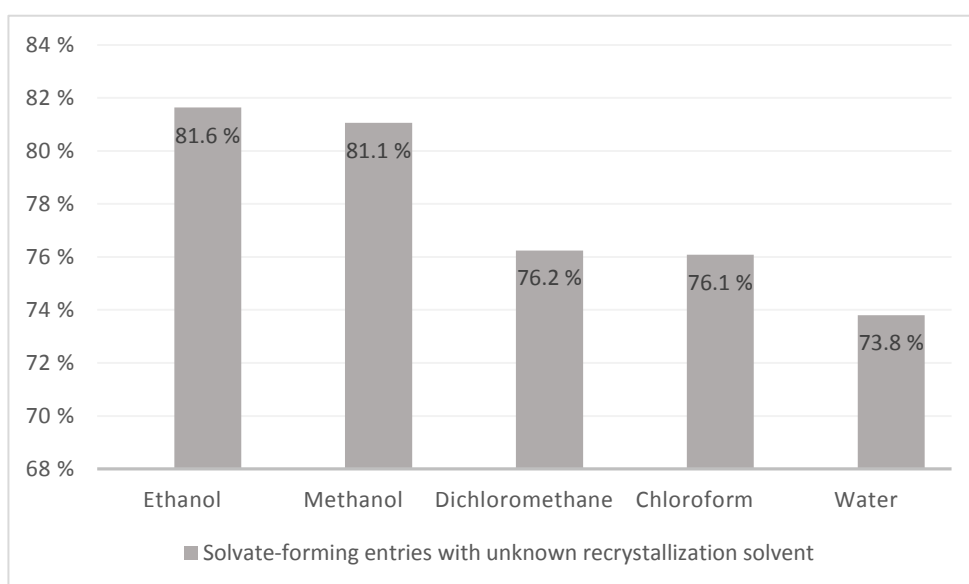


Figure 4-3. Percentages of entries with unknown recrystallization solvent among the solvate-forming entries in each solvent dataset.

The large percentage of empty recrystallization solvent entries raises the question whether to include these entries in the analysis or to exclude them. Since the solvent exists as a part of the structure in the solvate-forming group, this means that the solvent was used in the preparation of the sample. Solvate entries with unknown recrystallization solvents were

considered as part of the dataset. For the non-solvate-forming group, the recrystallization solvent cannot be known without having it mentioned in the database. For this reason, we excluded any non-solvate molecule with an unknown recrystallization solvent. This issue of an unknown recrystallization solvent did not happen in the non-solvate groups as the string of the solvent (for example “dichloromethane”) was searched for.

(3) Some solvate entries had a solvent in the crystal structure that was not included as a recrystallization solvent in the database. An example is the CSD reference code: ZEQPUI entry.⁹ This entry is an ethanol solvate in which the recrystallization solvent was recorded as a mixture of petroleum ether and ethyl acetate. Few entries (less than 10) of this type were present in each dataset. Since the solvent cannot just show up in the crystal, it must have been used although it might not be recorded. For this reason, these entries were accepted in the solvate dataset. They were rejected in the non-solvate datasets.

(4) Another issue that affected the accuracy of the solvate-forming lists was the inconsistency in the solvents nomenclature in the CSD, which made the collection of a complete dataset harder. For example, ethanol was recorded in the database under different names, including alcohol, ethyl alcohol, abs. alcohol, 96 % alcohol. The proportion of structures crystallized from ethanol that were saved using these names was below 0.5 %. In water datasets on the other hand, more than 13 % of the total non-hydrates had their recrystallization solvent recorded as “aqueous” instead of “water”, which can cause a significant change in the information obtained. These different solvent names were searched and the results were considered part of the dataset.

(5) It is important to mention that by applying the mentioned criteria in ConQuest, hydrate and solvate structures of different stoichiometric ratios (disolvates/ dihydrates/ hemisolvates etc.) were all considered to be a hydrate or a solvate form. An example of that is the carbamazepine dihydrate (CSD reference code: FEFNOT02)¹⁰ which contains two water moieties, yet was still

present in the datasets of hydrate although the criterion of the number of chemical units was set to two. This is because the search option used in ConQuest treats every molecule in the asymmetric unit as one chemical entity regardless of the number of times that this unit is present in the unit cell. On the other hand, entries that contained molecules with more than one type of solvent in the crystal structure were excluded from this search because they are considered to have three chemical units. An example is 2-ethoxy-1,3-bis(3-(trifluoromethyl)phenyl)-1,3-dihydro-1H-imidazo(4,5-b)quinoxaline methanol ethanol solvate, (CSD reference code: CAWREY).¹¹ This entry contains ethanol and methanol moieties as can be seen in Figure 4-4.

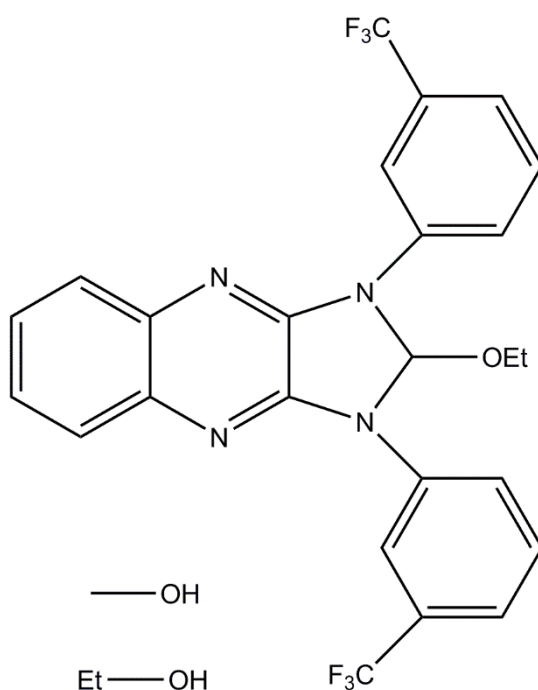


Figure 4-4. The CSD entry CAWREY showing more than one solvent in the crystal structure.

Entries with more than one type of solvent in the structure were excluded in order to reduce the number of variables in the investigation and to keep it as simple as possible. The count of the solvate and non-solvate forming entries after corrections are shown in Table 4-2.

Table 4-2. The count of solvate and non-solvate entries after refinement

	NS	S	Total
Ethanol	6,745	855	7,600
Methanol	4,065	1,942	6,007
Dichloromethane	1,520	1,767	3,287
Chloroform	1,449	1,676	3,125
Water	443	7,086	7,529
Total	14,222	13,326	27,548

4.2.4 Further partitioning of the solvate group

Including the solvate entries with unknown recrystallization solvent to the solvate-forming entries gave the question whether the entries recrystallized from a known source would show a different behaviour from the ones with no recrystallization solvent mentioned. It would also be advantageous to know if entries recrystallized from one solvent would show a different behaviour from the entries recrystallized from a combination of solvents. Note that the latter comparison is valid only if the two groups had a solvent in common. For these purposes, the data of each solvent was further partitioned as illustrated in Figure 4-5.

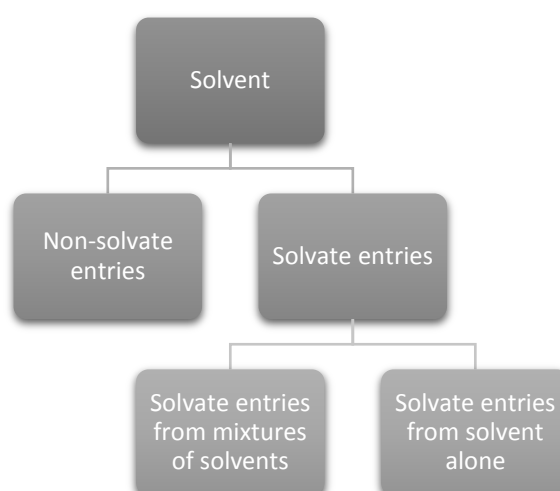


Figure 4-5. An illustration of the four individual groups of each solvent.

At this point, the solvate data consisted of three groups, the main group that includes all solvate entries, a sub-group that consists of entries recrystallized from a mixture of solvents and a sub-group that consists of entries recrystallized from the solvent of interest only. The non-solvate group was not partitioned; simply because the group did not have entries with blank recrystallization solvent tab in the database. The detailed number of entries in each of the solvate group and sub-groups are given in Figure 4-6.

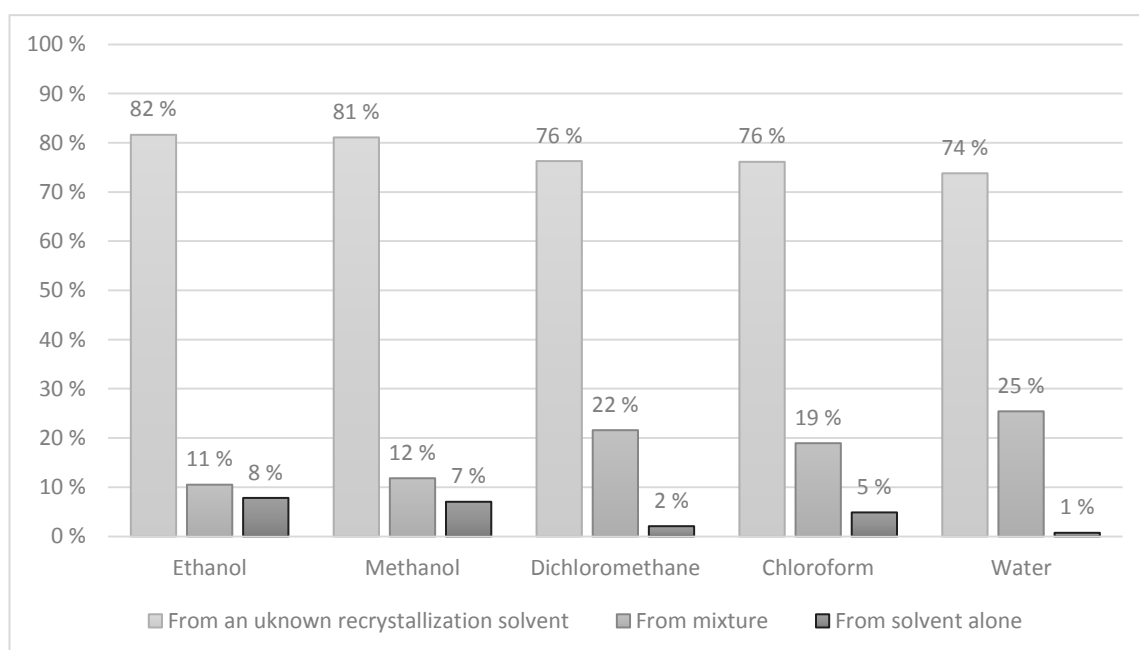


Figure 4-6. The number of entries that were recrystallized from the solvent of interest and the number of entries recrystallized from a mixture as obtained from the database. Note: numbers were rounded to the nearest integer this is why they do not all sum precisely to 100 %.

Based on the partitioning seen in Figure 4-5, a total of 20 groups (4 groups per solvent) were exported from the database as (gcd) lists. These are text files that include the CSD reference codes of the entries. They can be read in ConQuest to recall the hits from the database. After that, the entries were finally exported in the (mol2) file format, which contains the structural information and is suitable for descriptor calculation.

4.2.5 Preparing the data for descriptor calculation

The purpose of calculating molecular descriptors is to be able to represent the molecular features, such as the size of a molecule in a numerical value and later, compare those using statistical methods. The Dragon software was used for the molecular descriptor calculation. As the name “molecular descriptor” implies, Dragon strictly works with molecules. That means any entries that have more than one molecule in the asymmetric unit would give an error and no descriptors for this entry would be calculated. Based on that, all solvate entries are going to show an error. Non-solvate entries with more than one molecule in the asymmetric unit are also going to fail the descriptor calculation. Moreover, any entries with disorders would also cause an error. In order to resolve this issue, every extracted structure had to go under a “splitting step”. This means the mol2 files were processed to obtain one molecule in asymmetric unit before the descriptor calculation can take place. An example of each case above is shown in Figure 4-7.

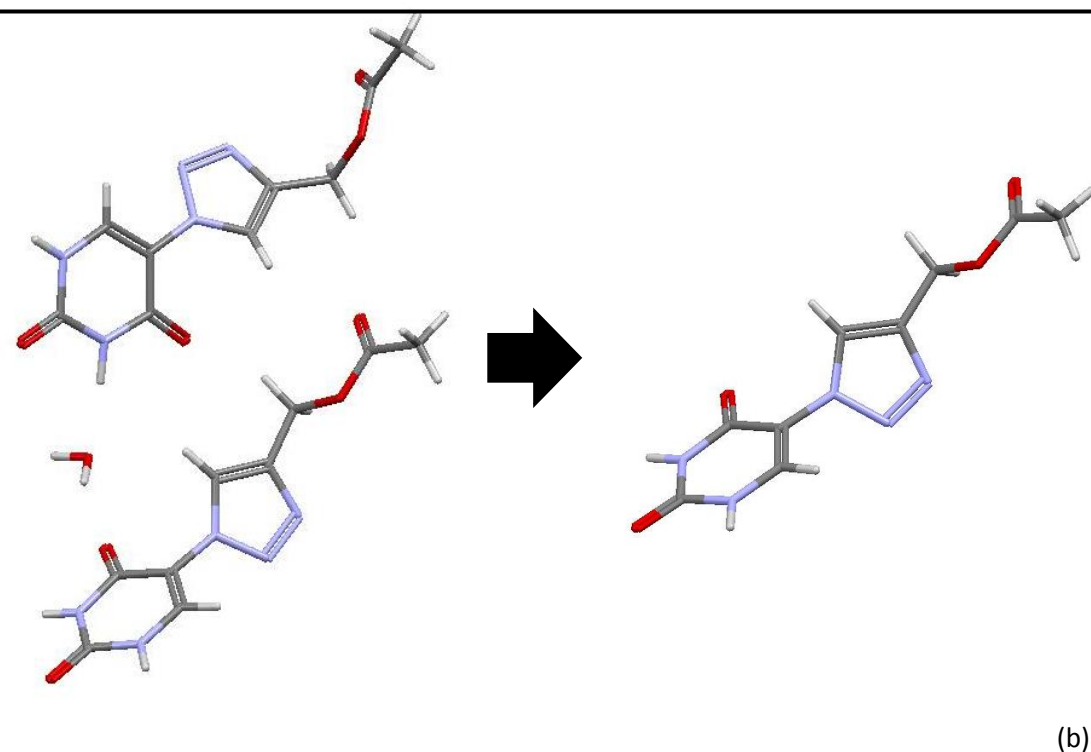
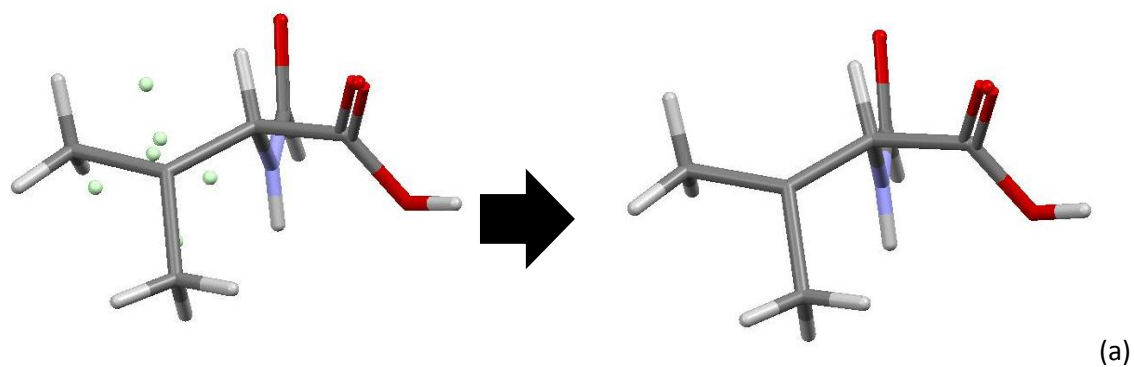


Figure 4-7. The splitting step of the AFEYOA¹² (a) and the COHLOC¹³ (b) entries.

The splitting step was performed using in-house software, developed by my PhD advisor, Dr. László Fábián (written in Perl language). This small program splits a mol2 file into a number of mol2 files equal to the number of molecules in the asymmetric unit. Each of the newly generated mol2 files has one molecule in it with the first file generated always having the molecule with the largest molecular weight.

4.3 Descriptor calculation

After all non-solvate and solvate molecules were extracted, refined and split, the 20 lists of entries were fed into Dragon, where 4,885 molecular descriptors were calculated per molecule. 20 tables, corresponding to the 20 lists of molecules were obtained. Each row in these tables represents a molecule (identified by its CSD reference code), while each column represents a molecular descriptor. A sample of the Dragon software output can be seen in Table 4-3.

Table 4-3. A sample from the Dragon software's output. Specifically, the sample was taken from the dichloromethane non-solvate entries dataset. NA values represent an error in the calculation

Refcode	<i>MW</i>	<i>AMW</i>	<i>nH</i>	<i>nAT</i>	<i>nSK</i>	<i>nHAcc</i>	<i>qpmax</i>
MUKBUQ	354.650	15.420	6	23	17	2	NA
FUGYAH	344.440	7.488	20	46	26	4	NA
BILCUU11	NA	NA	NA	NA	NA	NA	NA
LOVBON	212.280	9.230	8	23	15	1	NA
DUWBUT01	306.490	6.255	26	49	23	2	NA
DUWBUT02	306.490	6.2550	26	49	23	2	NA
QUPKIV	260.320	16.270	0	16	16	2	NA
WIZMAU	287.200	11.046	6	26	20	6	1

In this table, *MW* is the molecular weight, *AMW* is the average molecular weight, *nAT* is the number of atoms in the molecule, *nH* is the number of hydrogen atoms in the molecule, *nSK* is the number of the non-hydrogen atoms in the molecule, *nHAcc* is the number of hydrogen bond acceptors carbon atoms and *qpmax* is the maximum positive charge in each molecule.

As it can be seen in the Table 4-3 few errors occurred during the calculation of the molecular descriptors. Three types of molecules causing errors were identified. The cause of these errors and how they were treated is discussed in section 4.3.1.

4.3.1 Entries with calculation problems

(1) For some molecules, the calculation of all 4,885 descriptors failed although no errors, disorders or other problems could be seen in the structure/entry. An example is the entry S-(4-Chlorophenyl) (4-chlorophenyl)thiosulfonate, CSD reference code: BILCUU11.¹⁴ Molecules giving this type of error were not traced back and fixed as this would be a time-consuming procedure. Alternatively, these faulty molecules were omitted from the datasets.

(2) Another example of molecules that caused trouble during descriptor calculation are molecules with no hydrogen atoms. These were mostly old entries with no H coordinates determined. This causes mistakes in the calculation of some descriptors. The simplest example on a descriptor that can be miscalculated due to this missing information is the nH descriptor, which counts of the number of H atoms. Other descriptors such as nAT, the number of atoms in a molecule, were affected by this error. An example of a molecule showing this error is shown in Table 4-3, entry CSD reference code: QUPKIV,15 which is illustrated in Figure 4-8.

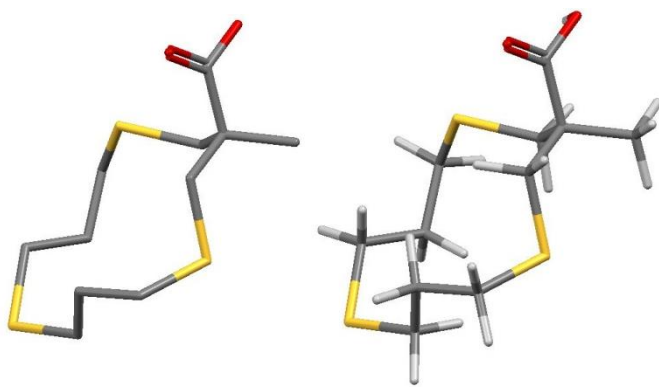


Figure 4-8. QUPKIV molecule and how it is recorded in the CSD (left) and the auto-edited structure by Mercury (right).

Molecules with no hydrogen atoms were also removed from all datasets using a script in R.

(3) Entries having polymorphs were prevalent in the extracted datasets. Since the original claim was to predict hydrate and solvate formation using the molecular graph (only 2-D

descriptors), polymorphs in the datasets cause a redundancy of the information. This can have large influence on the outcome of the data analysis. A script in R was developed to keep one of the polymorphs. The selection of a polymorph was based on keeping one entry with a unique molecular weight and the same first six letters of the CSD reference code.

The removal of faulty molecules, molecules with no hydrogens and polymorphs has reduced the number of molecules in the dataset. The new number of molecules in each dataset is shown in Table 4-4.

Table 4-4. The count of solvate and non-solvate entries left after removing the errors

Solvent	Number of structures	Solvates	Non-solvates
Ethanol	4,895	689	4,206
Methanol	4,366	1,518	2,848
Dichloromethane	2,761	1464	1,297
Chloroform	2,556	1,363	1,193
Water	4,432	4,128	304
Total	19,010	9,162	9,848

4.3.2 Descriptors with calculation problems

The mol2 files that were extracted from the CSD do not contain information about partial charges of molecules. For this reason, any charge-related descriptors showed NA values across the dataset. An exception was zwitterionic molecules, which had an overall charge of zero but contained formal charges in the molecule. An example of these descriptors is q_pmax, which is shown in Table 4-3. The WIZMAU16 molecule (for which the descriptors were calculated in Table 4-3) is shown in Figure 4-9.

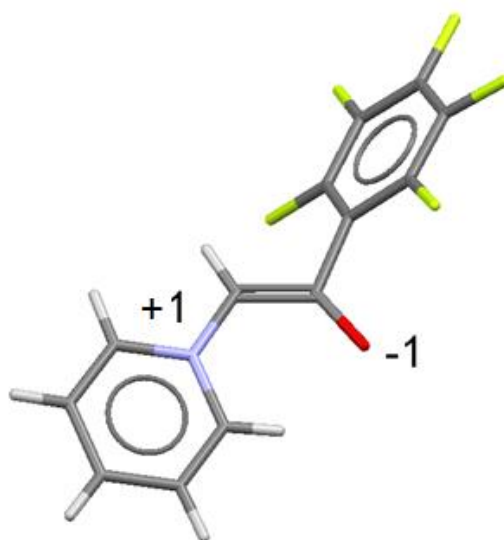


Figure 4-9. A zwitterionic molecule from the dichloromethane non-solvate dataset. The WIZMAU molecule.

These descriptors showed an NA value for more than 99 % of the data in each dataset. They were removed from the dataset because they were not meaningful at this point. The molecules that were zwitterionic were kept as part of the dataset.

4.4 Summary of data

Before starting the statistical analysis, it is a good idea to investigate the structure of the data that is going to be analysed. Every solvent had a different number of entries in its dataset. An illustration of the relative number of these entries is shown in Figure 4-10.

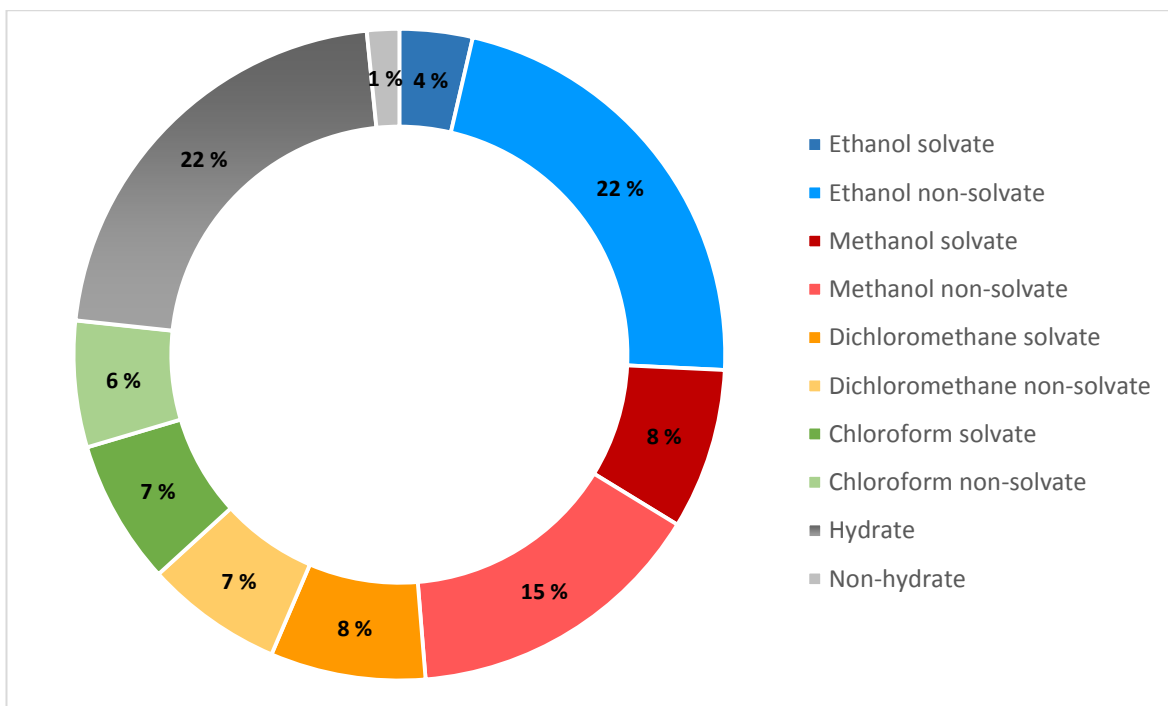


Figure 4-10. Percentage of solvate and non-solvate forming molecules in each solvent's dataset. Total number of entries is 19,010.

The number of entries in the ethanol set was the largest (26 %) followed by water (23 %), methanol (23 %), dichloromethane (15 %) and chloroform (13 %). In terms of solvate- and non-solvate groups, the hydrate and ethanol non-solvate were the largest among them. Ironically, the groups that corresponds to these two (the non-hydrate and ethanol solvate, respectively) were the smallest. Methanol has shown a more balanced dataset with a non-solvate group that is twice the size of the solvate group. Dichloromethane and chloroform have shown datasets that are almost perfectly balanced (nearly half-split of solvate and non-solvate).

After the structure of the data was seen, it was important to consider the range of data covered in these datasets. For example, what is the chemical space that is covered by these datasets? How hydrophilic are these compounds? Each solvent's dataset was plotted in terms of chemical space, Log P value, number of hydrogen bond acceptors and number of hydrogen bond donors, as illustrated in Figures 4-11 to 4-14. The column colours in these figures are just for a more clear illustration.

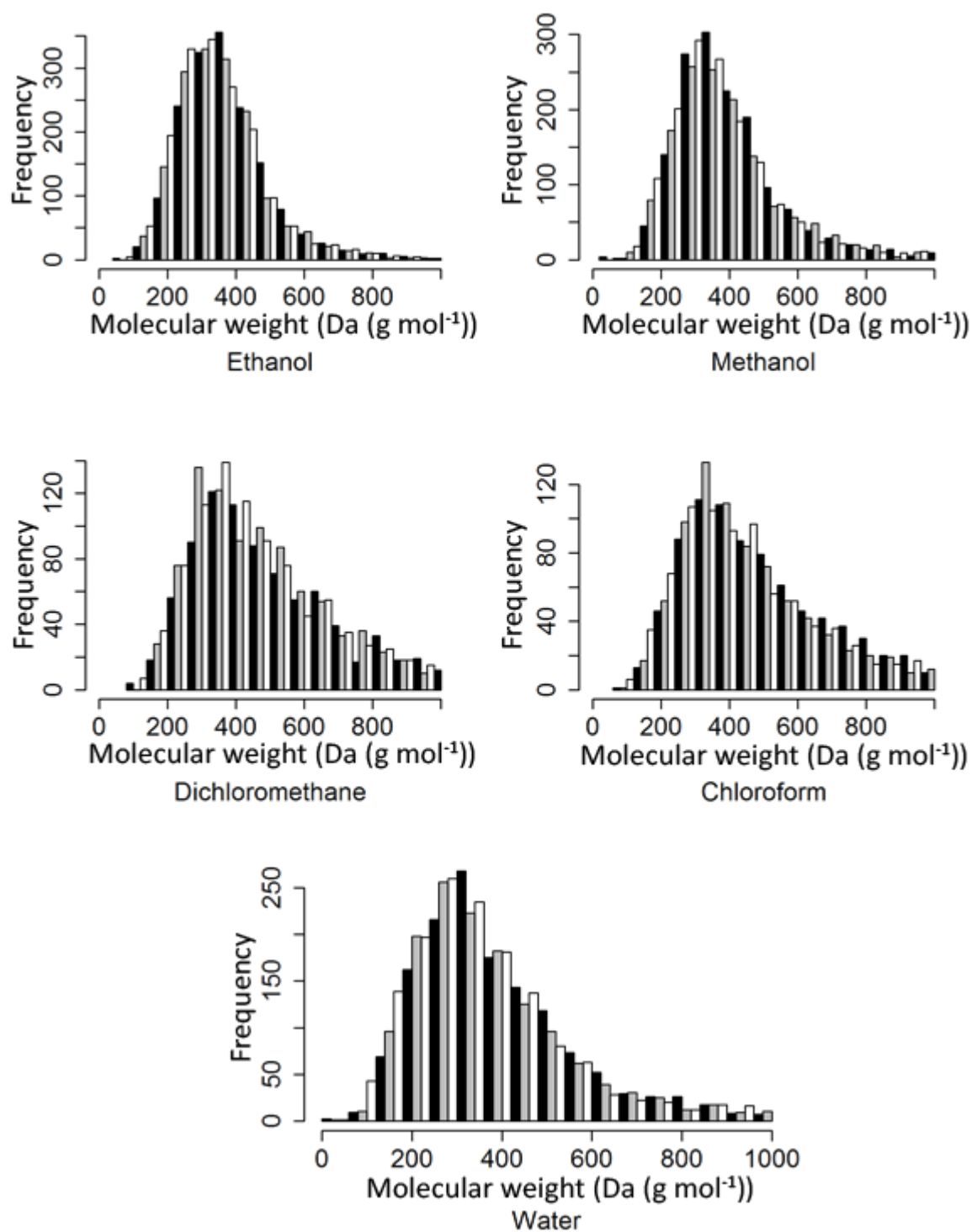


Figure 4-11. The chemical space covered by each solvent dataset (units in Da (g/mol)). Y axis shows the frequency of occurrences.

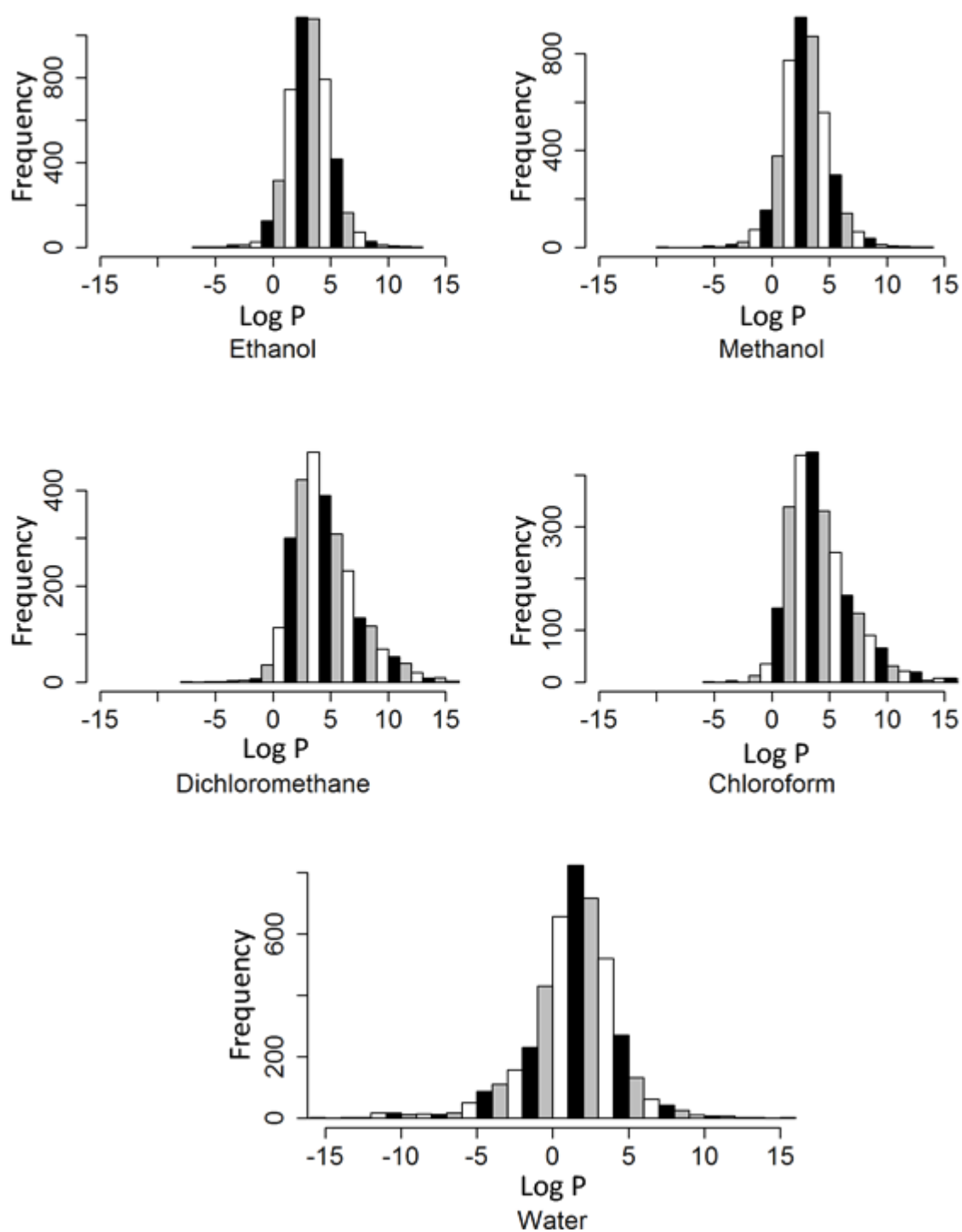


Figure 4-12. Log P values for each solvent dataset. Y axis shows the frequency of occurrences.

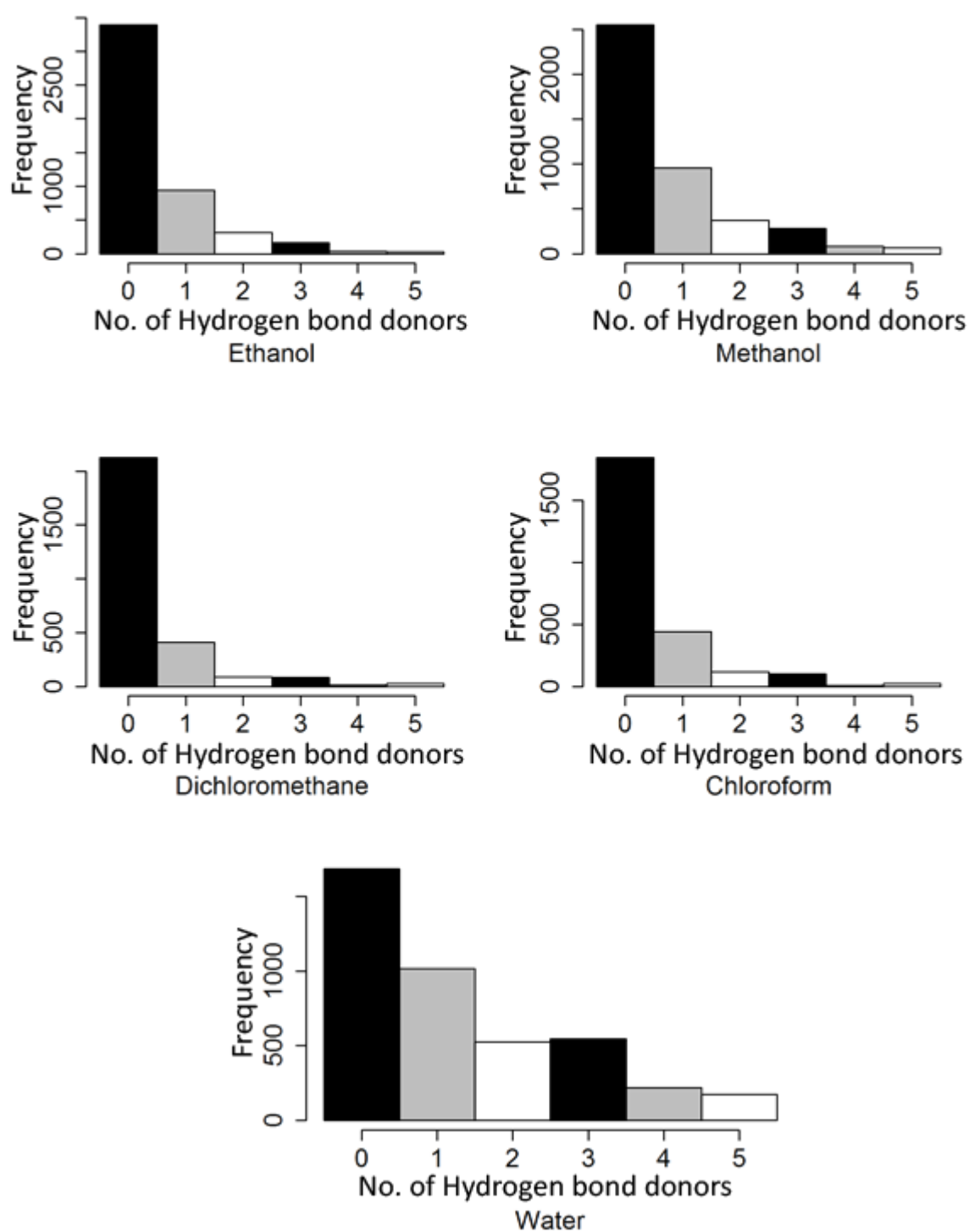


Figure 4-13. The number of hydrogen bond donors in each solvent's dataset. Y axis shows the frequency of occurrences.

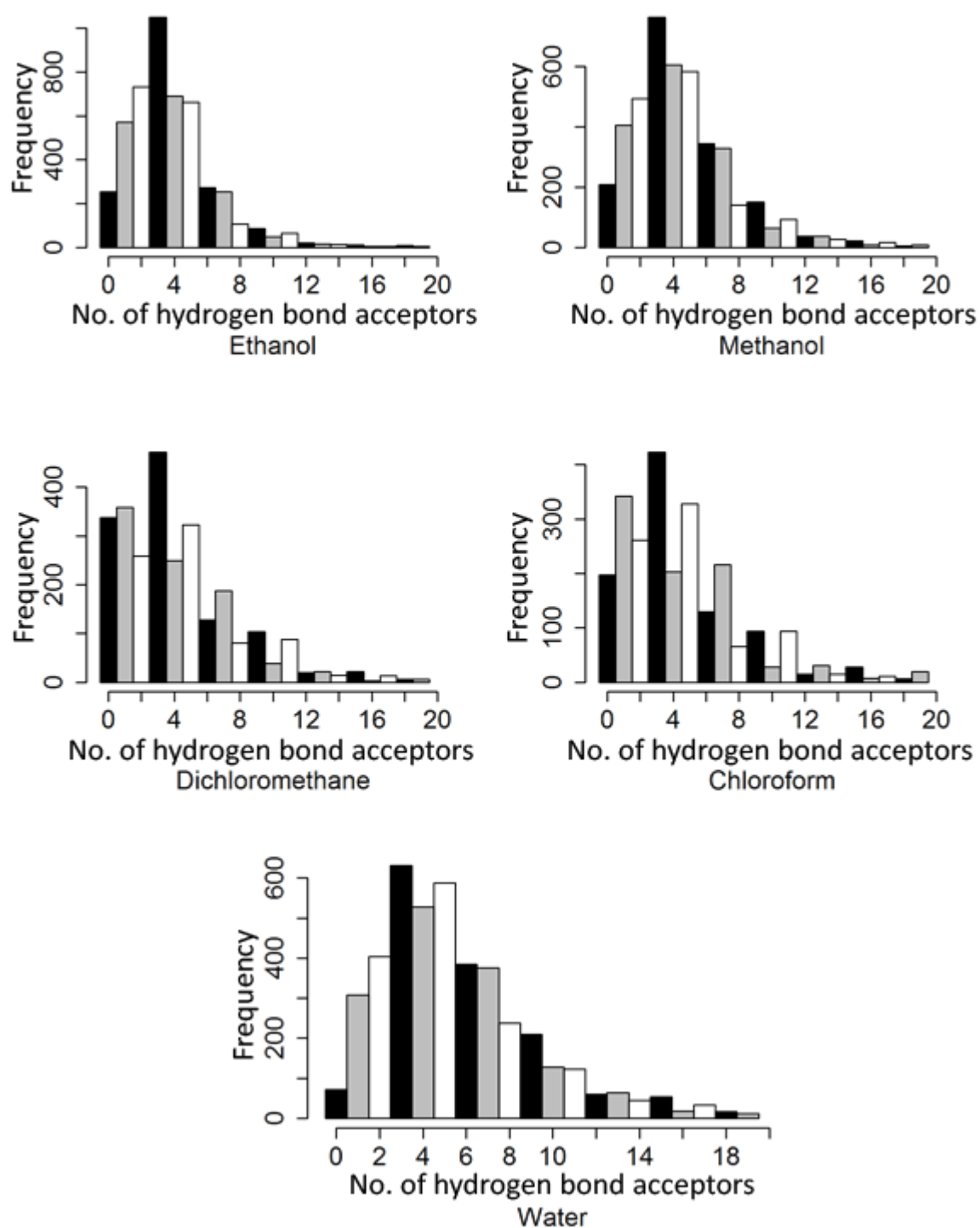


Figure 4-14. The number of hydrogen bond acceptors in each solvent's dataset. Y axis shows the frequency of occurrences.

As can be seen in Figure 4-11 the molecular weight of the entries in different datasets was similar, where most of the data was concentrated in the region of 100-600 Da (g/mol). The

chlorinated solvents have shown a larger number of high molecular weight entries compared to the other solvents. Water on the other hand has shown a slight shift towards lower molecular weight.

With exception of water, the Log P values of most entries ranged between -1 and 6 (Figure 4-12), showing a more hydrophobic nature of data for these 4 solvents. The chlorinated solvents have shown a tail towards higher values. The Log P values in most water entries ranged between -5 and 6, showing a more balanced hydrophilicity in its dataset.

As illustrated by Figure 4-13, most of the entries in the dataset possessed no hydrogen bond donors. In ethanol and methanol, the non-solvate entries largely outnumber the solvate ones, therefore a majority of entries with no hydrogen bond donors is a logical outcome. Chlorinated solvents are unable to form hydrogen bonds, so a majority of entries with no hydrogen bond donors can also be understood. On the hand, over 90 % of the water dataset were solvate entries, yet it had a majority of entries with no hydrogen bond donors. This is a rather interesting observation, as hydrogen bonding is reportedly one of the most important features that make hydrates more common than other solvate forms. For the number of hydrogen bond acceptors, solvents with hydrogen bond ability (ethanol, methanol and water) have shown a normal-like distribution (Figure 4-14) with the largest frequency being entries with 3 acceptors. Chlorinated solvents on the other hand, have shown fuzzy distribution, which does not show a clear pattern.

4.5 References

1. Görbitz CH, Hersleth H-P. On the inclusion of solvent molecules in the crystal structures of organic compounds. *Acta Crystallographica Section B: Structural Science*. 2000;56(3):526-34.
2. Grodowska K, Parczewski A. Organic solvents in the pharmaceutical industry. *Acta Poloniae Pharmaceutica - Drug Research*. 2010;67(1):3-12.
3. Clarke HD, Arora KK, Bass H, Kavuru P, Ong TT, Pujari T, Wojtas L, Zaworotko MJ. Structure– stability relationships in cocrystal hydrates: Does the promiscuity of water make crystalline hydrates the nemesis of crystal engineering?. *Crystal Growth & Design*. 2010 Mar 30;10(5):2152-67.
4. Kimura ET, Ebert DM, Dodge PW. Acute toxicity and limits of solvent residue for sixteen organic solvents. *Toxicology and Applied Pharmacology*. 1971;19(4):699-704.
5. World Health Organisation: Chapter 5.7: Dichloromethane. 2000.
6. Foye WO, Lemke TL, Williams DA. Foye's principles of medicinal chemistry. Sixth edition ed: Lippincott Williams & Wilkins; 2008.
7. Desiraju GR. Hydration in organic crystals: prediction from molecular structure. *Journal of the Chemical Society, Chemical Communications*. 1991(6):426-8.
8. Aznar R, Grabulosa A, Mannu A, Muller G, Sainz D, Moreno V, et al. [RuCl₂(η⁶-p-cymene)(P^{*})] and [RuCl₂(κ-P^{*}-η⁶-arene)] Complexes Containing P-Stereogenic Phosphines. Activity in Transfer Hydrogenation and Interactions with DNA. *Organometallics*. 2013;32(8):2344-62.
9. Zhang X, Han JJ, Huang M, Zhang GY. Synthesis, crystal structures and antibacterial activities of schiff base ligand and its cobalt(II) complex. *Russian Journal of Coordination Chemistry*. 2012;38(8):560-6.

10. Harris RK, Ghi PY, Puschmann H, Apperley DC, Griesser UJ, Hammond RB, et al. Structural Studies of the Polymorphs of Carbamazepine, Its Dihydrate, and Two Solvates. *Organic Process Research & Development*. 2005;9(6):902-10.
11. Schramm F, Walther D, Görls H, Käßlinger C, Beckert R. Trifluoromethylaryl-substituted quinoxalines: Unusual ruthenium-amidinate complexes and their suitability for annellation reactions. *Zeitschrift für Naturforschung B*. 2005;60(8):843-52.
12. Boyle GA, Govender T, Kruger HG, Maguire GEM, Negus TK, Rademeyer M. (S)-(+)-2-Formylamino-3-methylbutanoic acid. *Acta Crystallographica Section E*. 2007;63(9):o3912.
13. Korunda S, Kristafor S, Cetina M, Raic-Malic S. Conjugates of 1,2,3-Triazoles and Acyclic Pyrimidine Nucleoside Analogues: Syntheses and X-ray Crystallographic Studies. *Current Organic Chemistry*. 2013;17(10):1114-24.
14. Bodrikov IV, Nikitina NV, Subbotin AY. Oxidative dimerization of sulfenyl chlorides into thiosulfonates under the action of hexamethylphosphoramide. *Doklady Chemistry (Translation of the chemistry section of Doklady Akademii Nauk)*. 2011;436(1):1-4.
15. Setzer WN, Liou S-Y, Easterling GE, Simmons RC, Gullion LM, Meehan EJ, et al. Synthesis and structural studies of hydrophilic mesocyclic trithioethers. *Heteroatom Chemistry*. 1998;9(2):123-8.
16. Yamada S, Ohta E. (1-Pyridinio)perfluorophenacylide: a new stable pyridinium ylide in the enol form. *Acta Crystallographica Section C*. 2008;64(4):o230-o2.

Chapter 5: **Statistical analysis**

5.1 Overview

The think process of this chapter is outlined in Figure 5-1.

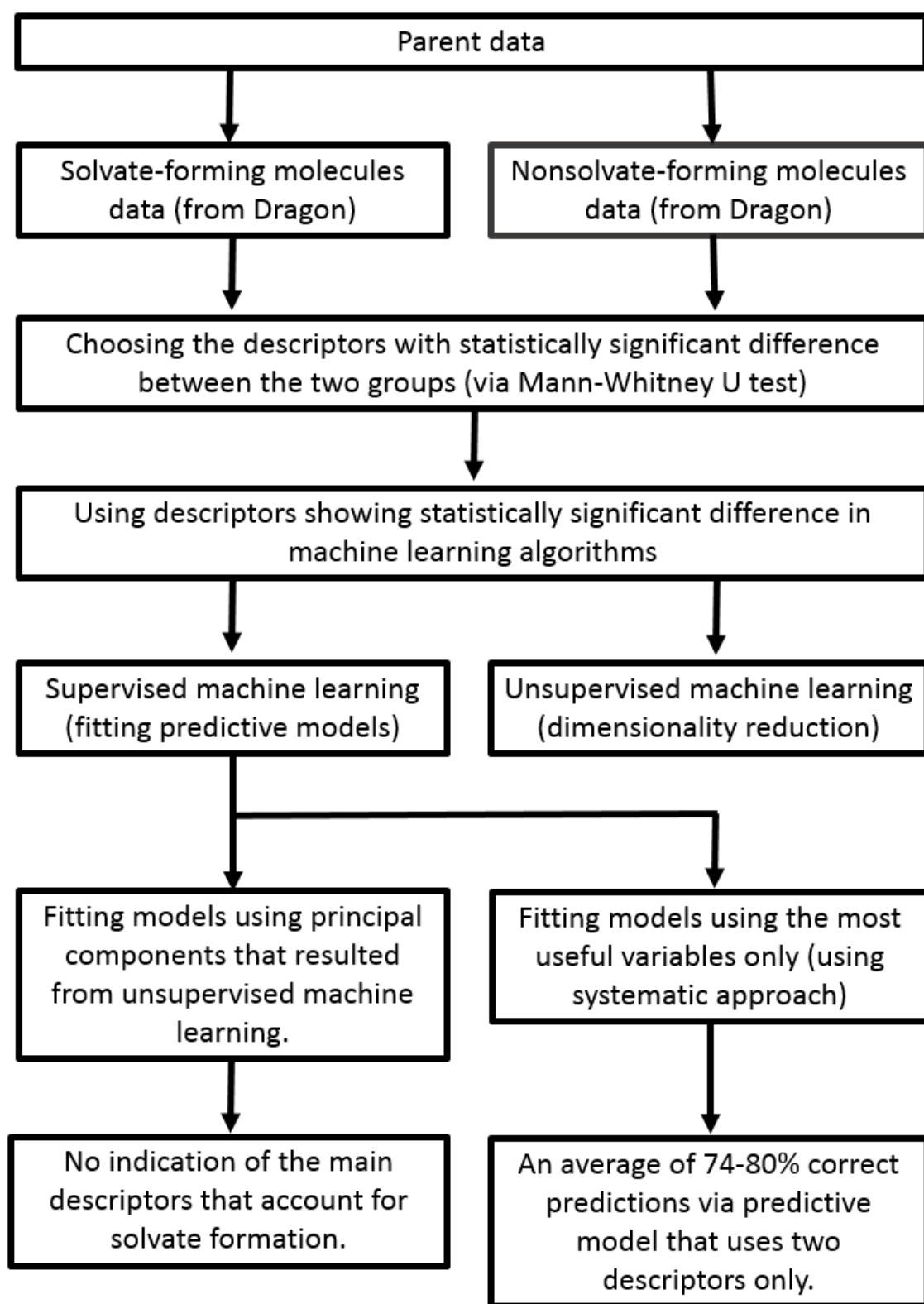


Figure 5-1. The thinking process throughout Chapter 5

5.2 Dimensionality reduction

The datasets obtained in Chapter 4 were now ready to start the analysis. Each dataset was described by about 5000 variables. The question here would be: are all these descriptors useful for predicting solvate formation? In order to answer this question, a significance test between each pair of solvate and non-solvate dataset was conducted. Such test would be able to tell which descriptors showed difference between the two groups, indicating the descriptors that are useful for solvate prediction. Ideally, it would also quantify that difference, indicating which variables show the largest discriminating ability between the two groups of interest. The Wilcoxon rank-sum test (also known as Mann–Whitney U test) was chosen to carry out this comparison.

5.2.1 The Wilcoxon rank sum test

This test works by testing a null hypothesis that the two samples being tested come from the same population. In this study, the null hypothesis would be that the solvate and the non-solvate groups come from the same population. In this comparison, the p-value was set to 0.05. This value is the conventional value that has been used for significance testing by Fisher¹. It corresponds to an α (alpha) level of 5 %, which means there is a 5 % risk of omitting a good descriptor. On the other hand, this also means that there is 95 % chance that the rejected descriptor is not useful. Since the values of descriptors can take any value, a two-tailed test was necessary. Because this is a two-tailed test, any descriptor with a p value above 0.025 was rejected.

The non-parametric nature of this test means it doesn't have any underlying assumptions about the distribution of the parent data. This was important as the datasets of the five solvents are not guaranteed to have a normal distribution. Moreover, the dataset is described by the molecular descriptors, where each of them represents the data differently. This renders the data not feasible for optimal transformation of each descriptor. Another property that was

important in choosing the Wilcoxon rank-sum test was its ability to rank the data, therefore, it is not largely affected by outliers. For these reasons, the Wilcoxon test was suitable for the data in question. A deeper insight of the characteristics of this test are given in the literature review (Section 2.3.1).

At this point, it is useful to remember the structure of the data. It consists of 5 datasets corresponding to the 5 solvents being tested. Each of these 5 datasets is partitioned into 4 groups. Each group consists of a number of molecules that are described by 4885 descriptors. The general structure of the data was illustrated in section 4.2.4.

As mentioned earlier, the Wilcoxon rank-sum test compares two samples of a population, in terms of one variable. In this sense, two aspects had to be adjusted for each solvent's dataset in order to run the significance. These are the number of samples (i.e. groups) and the number of variables (i.e. descriptors). In order to adjust the number of samples, the test was applied to the two major groups only, that is; the group consisting of all non-solvate forming molecules (NS) and the solvate-forming molecules (S). The other groups, i.e. the solvate from a mixture of solvents (S-M) and the solvate entries from the solvent of interest alone (S-O), were just included in the visualization of the results. In order to compare one descriptor at a time, the test took place on a descriptor-by-descriptor basis. This means every descriptor from the non-solvate-forming molecules (NS) dataset was compared with the same descriptor from the corresponding solvate-forming molecules (S) dataset at a time. An automated loop was scripted in R to perform the analysis over the 4885 descriptors. The p-value for each comparison in the loop was recorded, and the descriptors that show a p-value larger than 0.025 were omitted from the dataset. Figure 5-2 shows an illustration of the comparison between the four groups in the ethanol dataset.

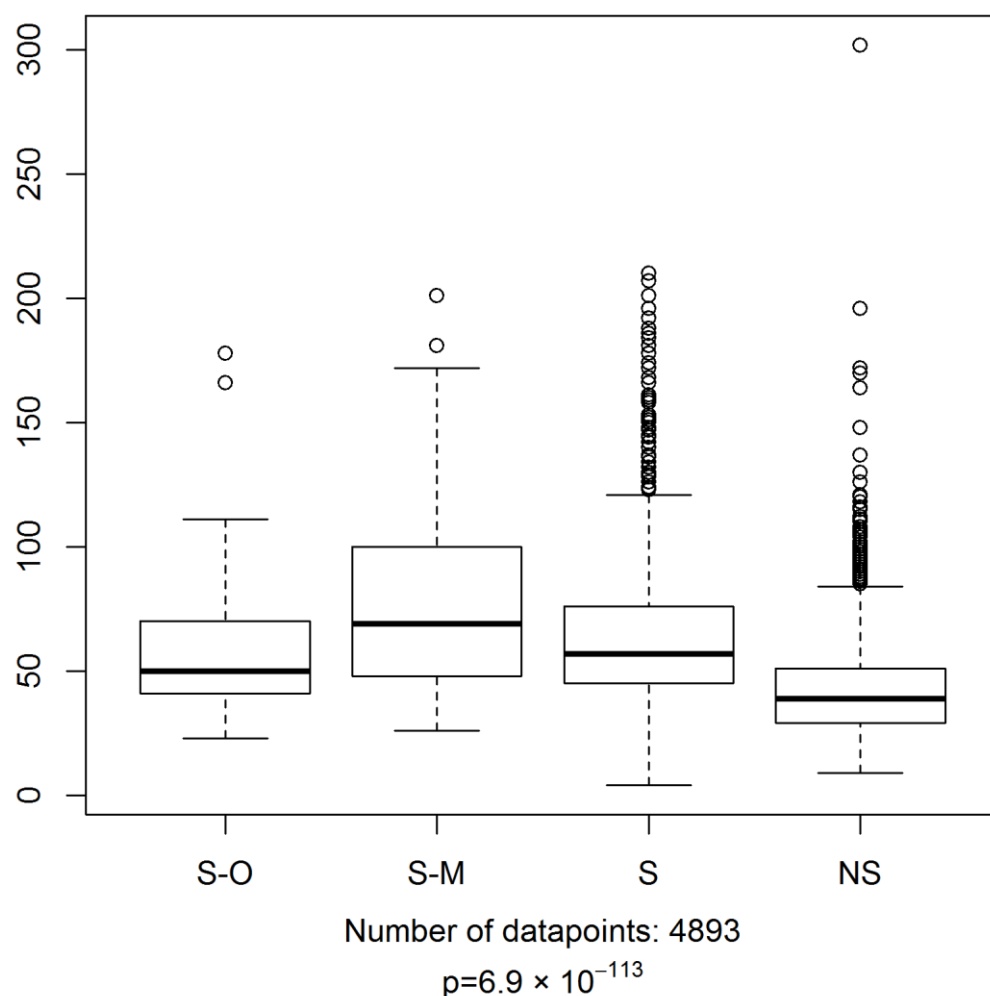


Figure 5-2. Boxplot of the nAT descriptor value in the ethanol dataset (Y axis). The circles represent the outliers in the dataset.

The boxplot shows a clear difference between the solvate-forming group and the non-solvate forming group in this descriptor, but the variation among the solvate sub-groups was minimal. This trend of the S-O and S-M groups to have similar descriptor values to the larger S group was observed for all descriptors. Despite the known effect of having a mixture of solvents on the recrystallization process (by changing solvent activity), this phenomenon cannot be observed using the descriptors and datasets included.

Descriptors that did not show a significant difference between the two groups were omitted from the datasets as these descriptors do not give information about what features might be important for solvate formation. This omission leaves us a smaller dataset to investigate for

each solvent. An example of a descriptor that did not show a significant difference between the two groups is shown in Figure 5-3.

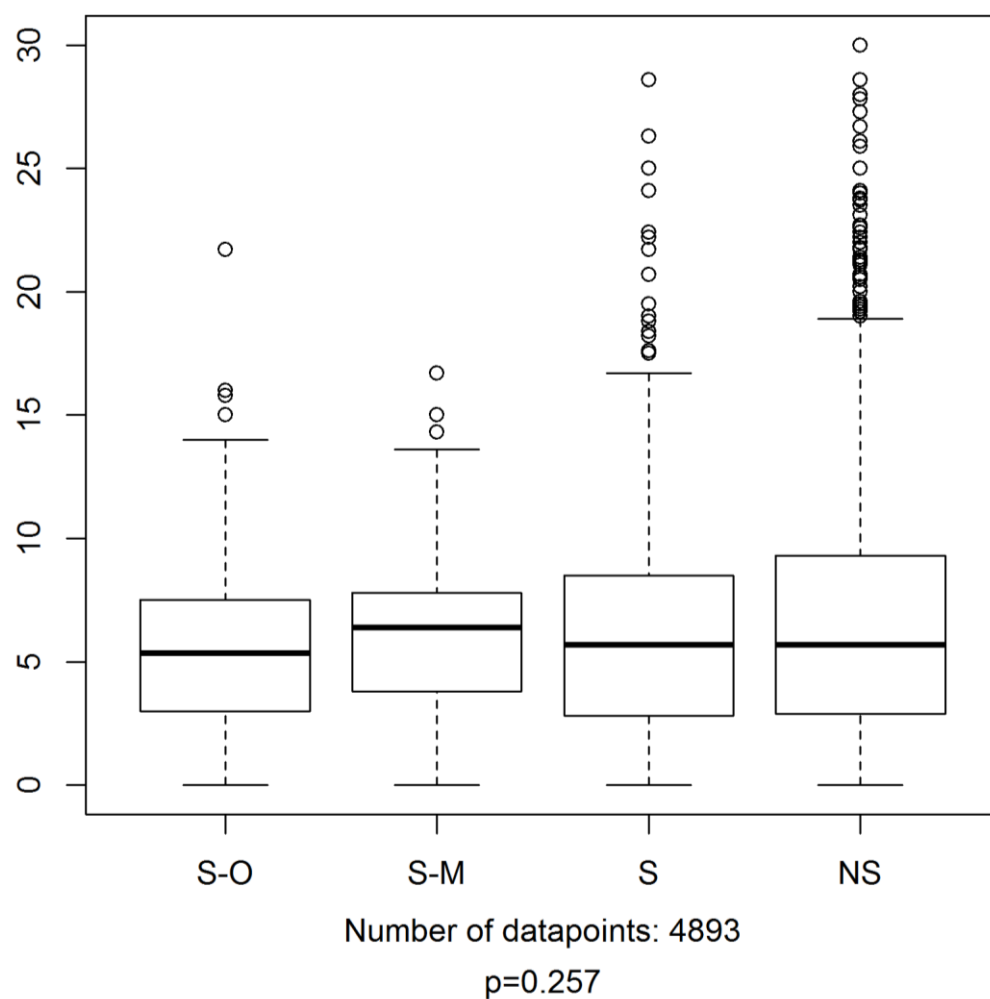


Figure 5-3. Insignificant difference between the groups in the O % descriptor. Such descriptors were omitted from the dataset.

Over 2850 out of 4885 descriptors turned out to have a significant difference between the solvate and the non-solvate group in each solvent's dataset. An illustration of the amount of remaining and omitted descriptors is shown in Figure 5-4.

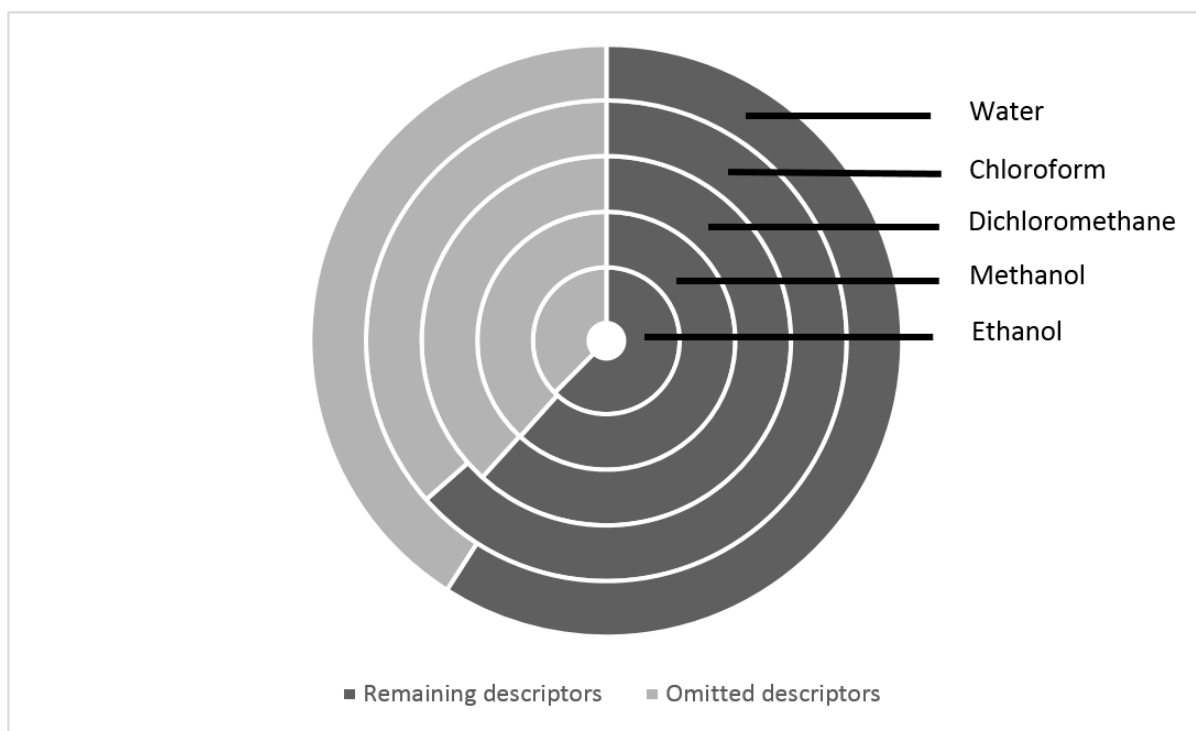


Figure 5-4. An illustration of the number of the variables that were omitted due to insignificance between the solvate-forming and the non-solvate forming groups. Total number of descriptors in each circle is 4885.

Although the test reduced the datasets to the descriptors with significant difference between the NS and the S groups, the reduction was not enough. 2850 descriptors in each dataset do not give a clear idea about the features that influence hydrate/solvate formation.

5.2.2 Principal component analysis

The main concern at this point was the large number of descriptors in each dataset. A method that can select the most meaningful descriptors in each dataset was required. This means a further reduction in the number of descriptors. It is important to remember that all these descriptors have shown significant difference between the groups and omitting them means the loss of part of the information they might contain. An unsupervised machine learning technique, principal component analysis (PCA) was the method chosen to solve the issue.

PCA reduces the dimensionality of a problem by creating new variables from existing descriptors. The PCA method has the advantage of combining the available variables (descriptors) rather than selecting the best ones in the dataset (See section 2.3.2.1). This

reduces the number of variables being studied without losing the information the descriptors might contain.

The principal component analysis procedure was applied to each solvent dataset individually. The principal component algorithm has given rise to a large number of principal components, equal to the number of datapoints in each dataset. For example, over 1300 PCs were given in the ethanol dataset.

The large number of PCs obtained were ordered by the amount of variance explained by each of them. The first principal component accounts for the highest variability in a dataset and the succeeding components account for lower variability. An illustration of the variance explained by the first 100 principal components in the ethanol dataset is shown in Figure 5-5.

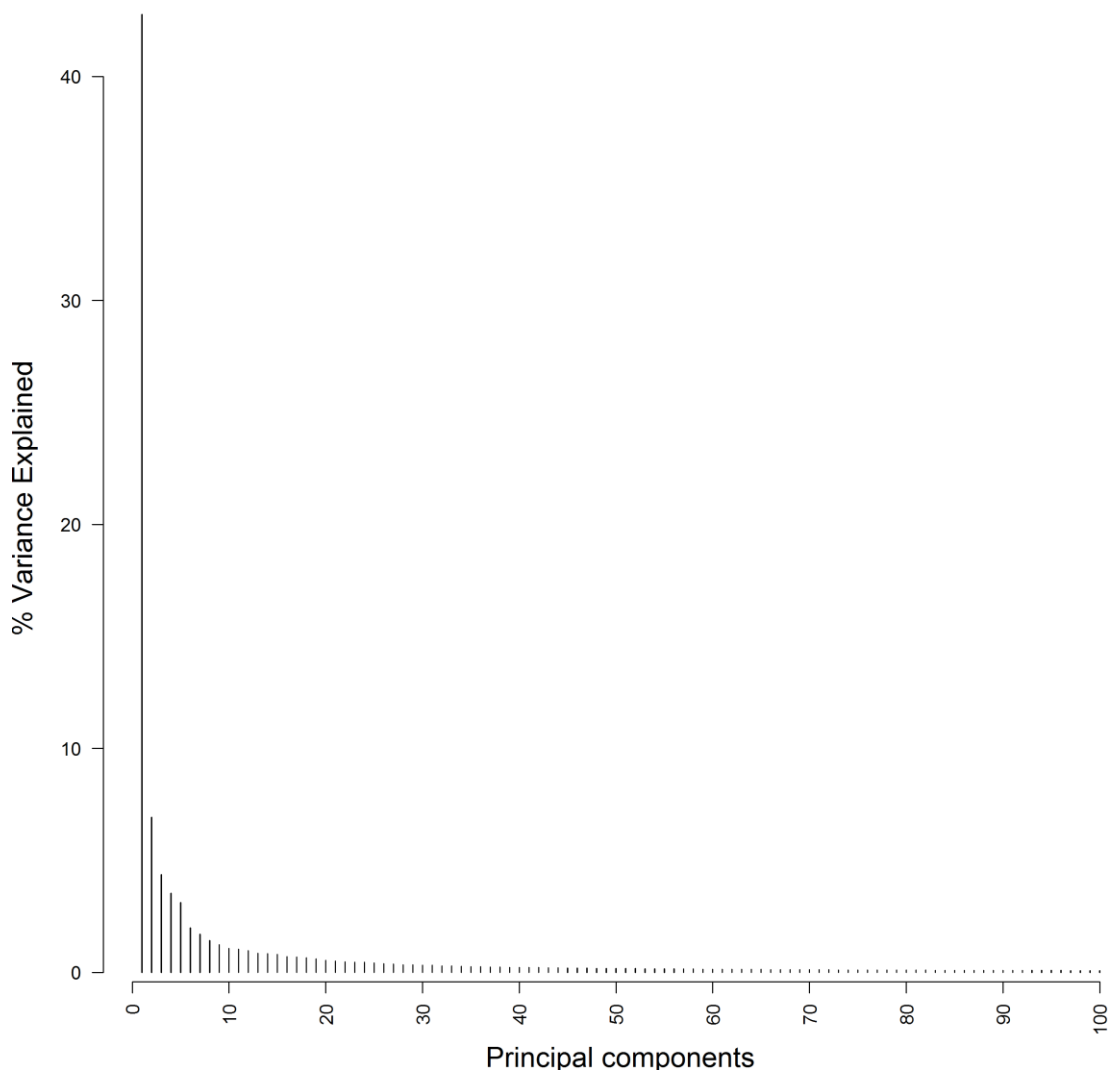


Figure 5-5. The percentage of Variance explained by the top 100 principal components in the ethanol dataset. Each bar represents a principal component, ordered.

The variability explained by principal components dramatically drops after the first few ones. For example, in the ethanol dataset, the top three principal components accounted for 54 %, while the top 5 accounted for 61 % and top ten principal components accounted for 68 % of the total variance. In order to see how well the PCA algorithm works in the different solvents datasets, the variance explained by the top three, five and ten principal components in each dataset was observed.

Although the percentage of variance explained by them isn't extremely high, the first 3 principal components seem a good choice to have an initial look at the data, as they explain a

variability equal to the rest of the variables in the dataset. Additionally, visualizing data in space with more than 3 variables would not be an easy task. For these reasons, only the first three principal components were selected. A scatterplot of the data points in terms of the principal components would show the spread of these points across the imaginary axes with the highest variability. It might as well show a trend of the solvate-forming and the non-solvate-forming molecules along the principal components. The spread of data points of the ethanol dataset in terms of PCA is shown in Figure 5-6.

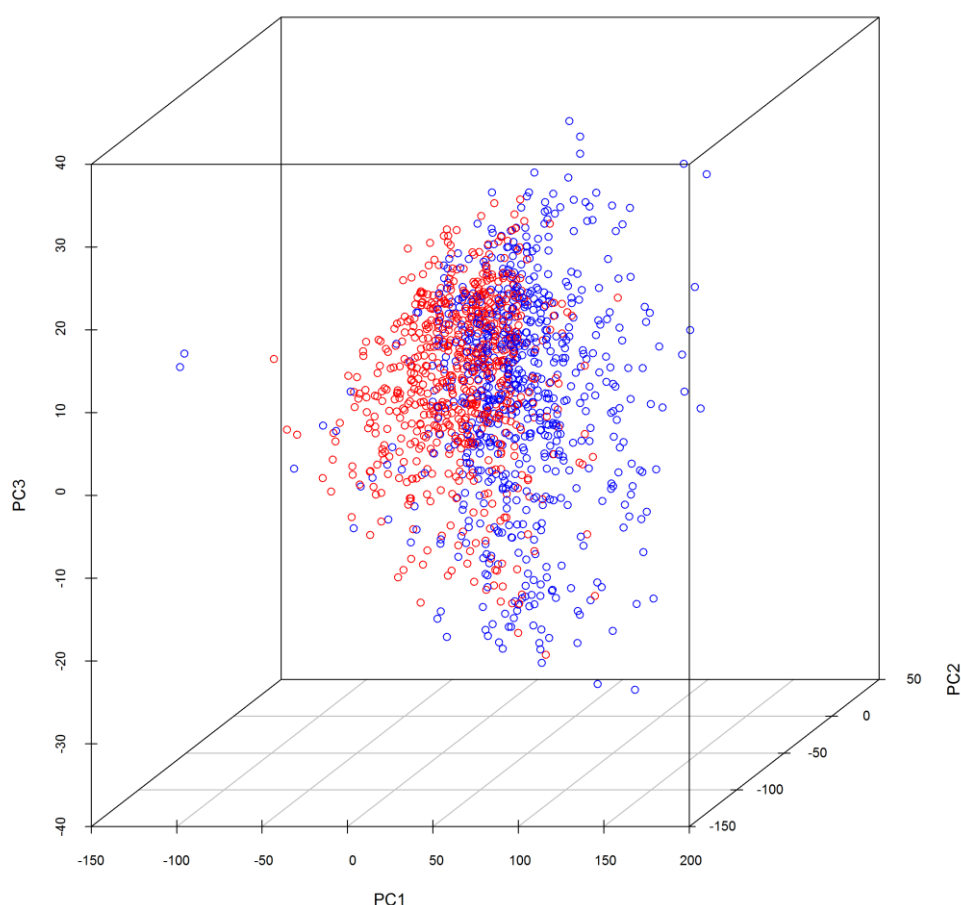


Figure 5-6. The ethanol data points in terms of the first three principal components. The solvate-forming molecules are shown in blue and the non-solvate forming molecules are shown in red.

A 3D plot from a 2D perspective isn't the best way to precisely see the effect of the principal components, but it can certainly show the combined effect of the three dimensions (principal components). For more clear representation of the spread of the data points in terms of the

first three principal components, pairs plot was used. An example of pairs plot of PC1, PC2 and PC3 of the same ethanol dataset is shown in Figure 5-7.

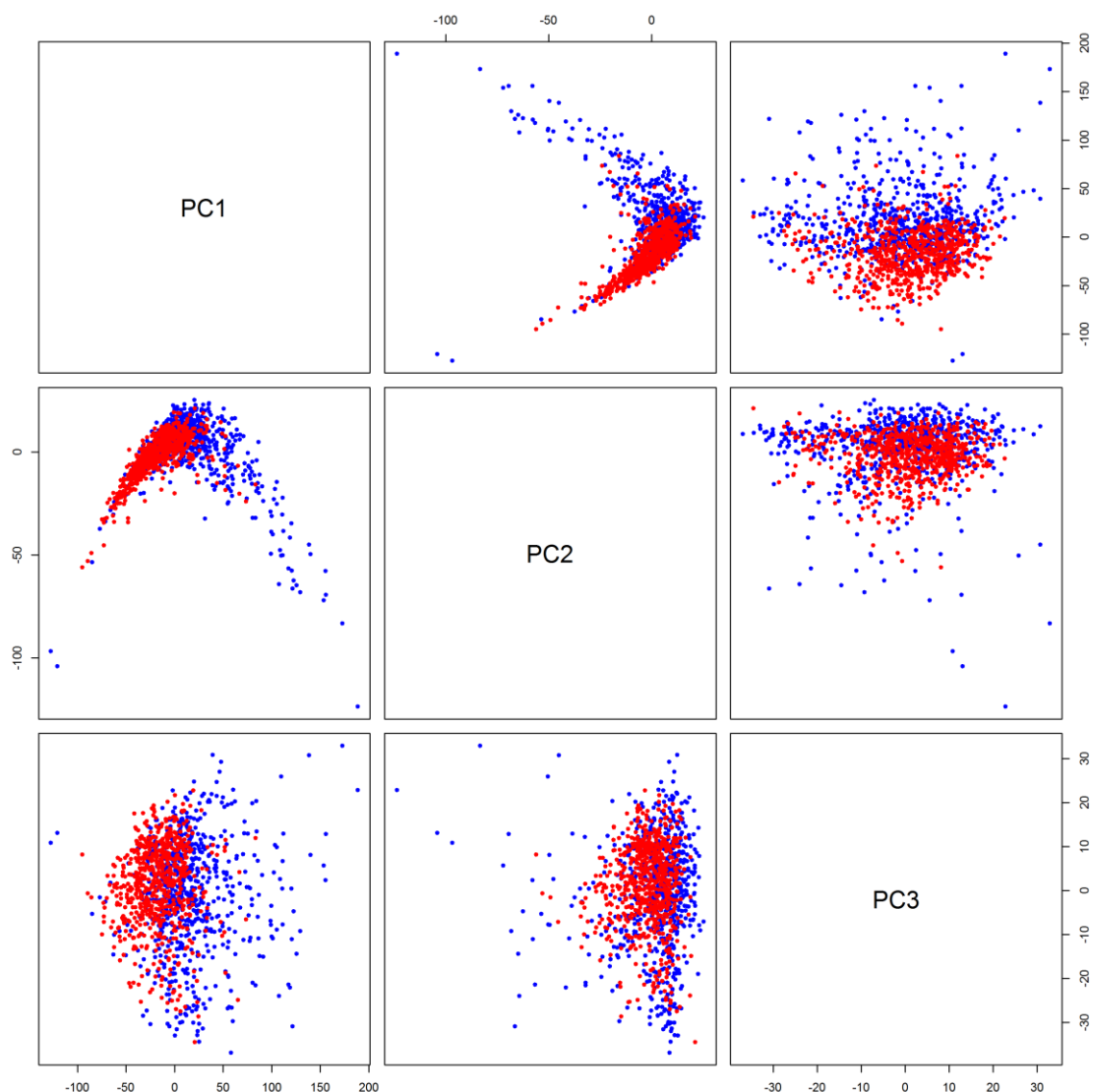


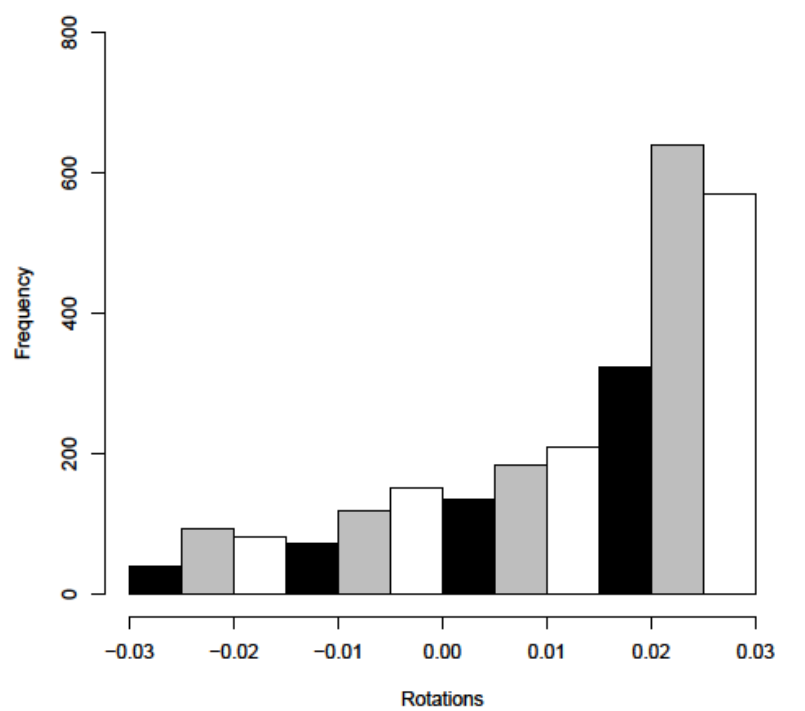
Figure 5-7. Pairs plot of ethanol data points in terms of PC1, PC2 and PC3. The solvate-forming molecules are shown in blue and the non-solvate forming molecules are shown in red.

The principal component analysis showed a good spread of the data points across the axes. Additionally, they showed a fairly good separation of the solvate-forming and non-solvate forming molecules along the top three PCs.

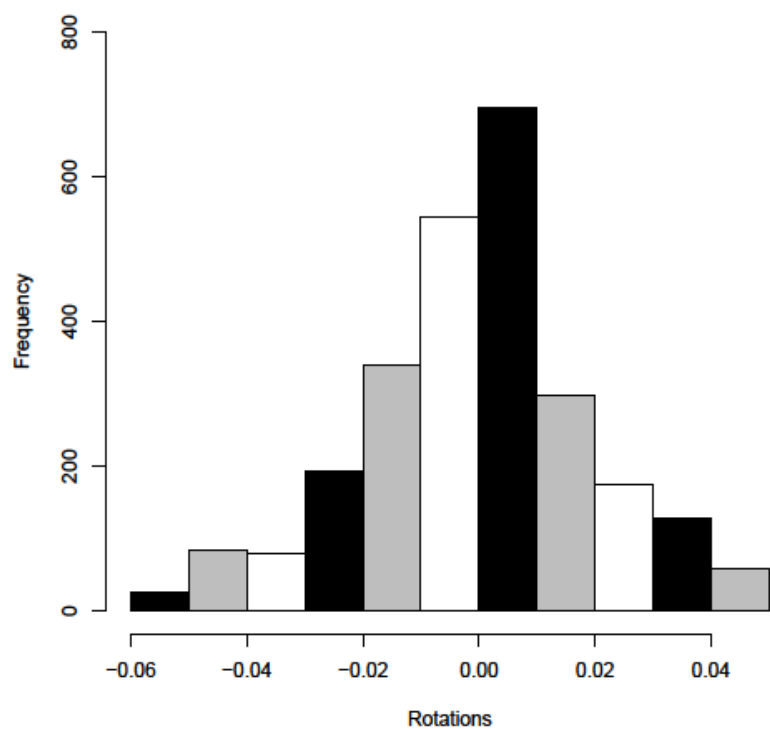
As it has been previously pointed out, the principal components are essentially combinations of the existing descriptors. Since these components showed some splitting of the data, the knowledge of which descriptors mainly account for these principal components might give an indication of the most important descriptors involved in solvate formation. In order to know

which descriptors affect the first three principal components, the loadings (rotations) of these principal components should be calculated. The loadings of a principal component show the contribution of each descriptor to this principal component ². This means that if a certain descriptor had a high rotation value of PC1 (e.g. 0.7), then it is an important determinant in the classification of the data points into solvate and non-solvate forming molecules.

A list of the rotation values of all descriptors was obtained in each solvent. An illustration of the distribution of the rotation values of PC1, PC2 and PC3 in the ethanol dataset is given in Figure 5-8.



(a)



(b)

Figure 5-8. Histograms of the rotation values of the first (a), second (b) and third (c) principal components in the ethanol's dataset. Total number of variables is 2647. The colours of the bars have no indication, they are used for a better illustration.

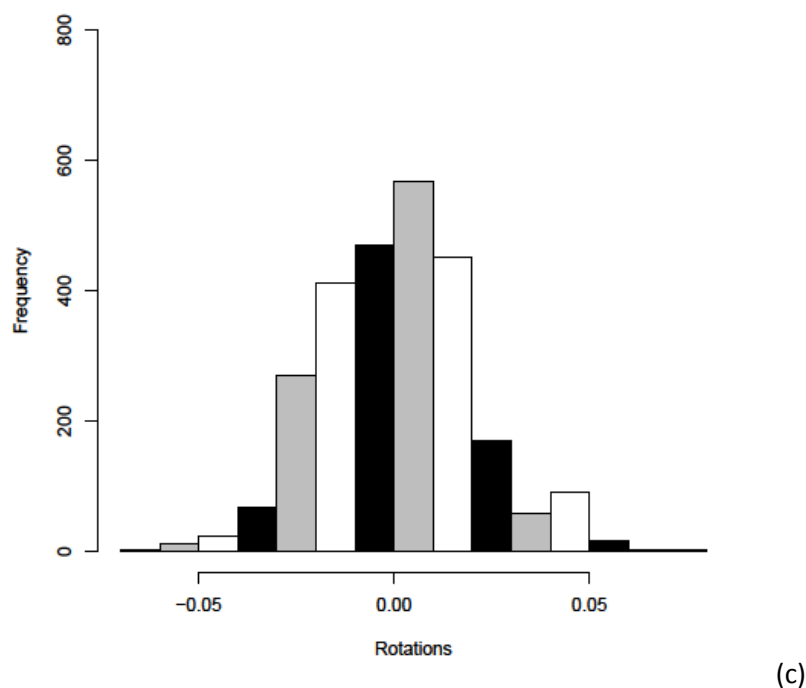


Figure 5-8. Histograms of the rotation values of the first (a), second (b) and third (c) principal components in the ethanol's dataset. Total number of variables is 2647. The colours of the bars have no indication, they are used for a better illustration.

Although the shape of the histogram shown in Figure 5-8 (a) looks promising (as there seem to be some high values at the positive end of the x-axis), the actual values on the x-axis are very small. Small rotation values were obtained for the first three principal components in all solvents datasets. The details of the maximum rotations of the first 3 PCs in each solvent are shown in Table 5-1.

Table 5-1. The maximum rotation value in the first three components in each solvent dataset

Solvent	Max rotation in PC1	Max rotation in PC2	Max rotation in PC3
Ethanol	0.028	0.047	0.068
Methanol	0.027	0.052	0.067
Dichloromethane	0.027	0.044	0.068
Chloroform	0.028	0.046	0.066
Water	0.027	0.039	0.076

This means that these principal components do not mainly depend on one or even on a few descriptors, which makes it hard to conclude which molecular features would affect the solvate formation with any solvent.

5.3 Supervised machine learning

Since PCA did not help in concluding what molecular features affect solvate formation, the direction of thinking has changed from dimensionality reduction to using a learning method in an attempt to find a pattern in the data. One special feature about this type of data is that it already has a known outcome, that is, each molecule is labelled as a solvate or a non-solvate. When this is the case, the supervised machine learning techniques can be used. (see section 2.3.2.2) Several machine learning algorithms are available. The question is which of these methods would be the most suitable in this case?

5.3.1 Selection of the machine learning algorithm

Tens of machine learning algorithms are available, and trying each for solving the solvation problem wouldn't be feasible. In order to select the most suitable machine learning algorithm, it is important to know the underlying concepts of the algorithms. Machine learning algorithms can be broadly classified into linear and non-linear classifiers. For example logistic regression (LR) is a linear classifier, while neural networks and support vector machines can perform as nonlinear classifiers. The ultimate way to know whether the problem of solvate formation is linear or non-linear (in terms of the molecular descriptors available) would be a preliminary testing of a linear and a non-linear classifier. Logistic regression was used as a representative of the former family while SVM (with an RBF kernel) was chosen as a representative of the latter family. The selection rationale of these methods was their simplicity and the previous evidence of their usefulness in classification problems. The RBF kernel is a reasonable choice for preliminary testing of SVM. It offers several advantages over other kernels, most notably

the small number of parameters that require adjustment compared to other kernels.³ The KNIME software was used to perform preliminary testing and comparison of these methods.⁴ This program provides a user-friendly interface and offers different machine learning algorithms, among which SVM and logistic regression are available. Note that the measure used for choosing the better method was the accuracy, which means the number of correctly predicted instances over the total number of instances. This is the method that is mostly used in supervised learning algorithm selection.⁵

Before the testing started, three “concerns” had to be adjusted, these are the number of descriptors, the number of molecules and the parameters of the machine learning methods, as explained by sections 5.3.2 to 5.2.4.

5.3.2 Choosing descriptors to decide the linearity of the problem

At this point, around 3,000 descriptors have shown to be useful to discriminate between the solvate and non-solvate groups in each dataset. It was not possible to use all these descriptors at once to fit a predictive model, helping to show the performance of LR and SVM. The large number of variables and the multicollinearity among the descriptors have caused errors and prevented the algorithm from converging. Alternatively, three variables were selected based on the p-value in the Wilcoxon rank sum test. The first variable to be selected was the one with the lowest p-value. The second variable shouldn't be selected directly to have the second lowest p-value. This is due to the possible high correlation between the first and the second two variables with the lowest p-values. Therefore, any variable showing a correlation above 0.5 to the first variable was removed from the dataset, and the variable with the lowest p-value among the remaining ones was selected. The third variable was selected in a similar manner to the second where any variable correlated more than 0.5 to the second variable was omitted from the dataset and the variable with the lowest p-value among the remaining ones was selected.

5.3.3 Equal size sampling

Datasets such as the ethanol and water datasets showed a high imbalance between the number of solvate and nonsolvate entries. If a logistic regression or a support vector machine model was to be fitted with any of the variables in these two datasets, the outcome of the prediction would be 100 % of the larger group. For example, a water model was fitted to the descriptor with the lowest p-value, which is the number of carbon atoms in the molecule (nC). The prediction results of this model are shown in Table 5-2.

Table 5-2. The confusion matrix and overall prediction accuracy of the water logistic regression model, fitted to the nC descriptor

	Predicted as non-hydrate	Predicted as hydrate
Non-hydrate	0	273
Hydrate	0	3714
Overall correct prediction	93 %	

The overall accuracy looks very promising, where 93 % of the data was correctly classified. On the other hand, the confusion matrix shows that none of the hydrate entries was predicted correctly. This leads to a completely useless model. A more reasonable representation would be using samples of equal sizes from both the solvate and the non-solvate datasets. This causes the total number of molecules (rows) in each dataset to drop. The total count of molecules was kept as large as possible by reducing the number of data-points in the group with the larger number just enough to meet the number of the data-points in the smaller group.

The random selection of the instances for equal size sampling could lead to the inaccuracy of conclusions, since not all molecules were used for fitting the models or even for testing them. This can be referred to as sampling error. To minimize this error, 3 equal size, random samples

were withdrawn from the complete dataset.⁶ Each of these equal size samples was treated as a separate dataset, where logistic regression and support vector machine models were fitted and the error rates were recorded. The steps of analysis are illustrated in Figure 5-9.

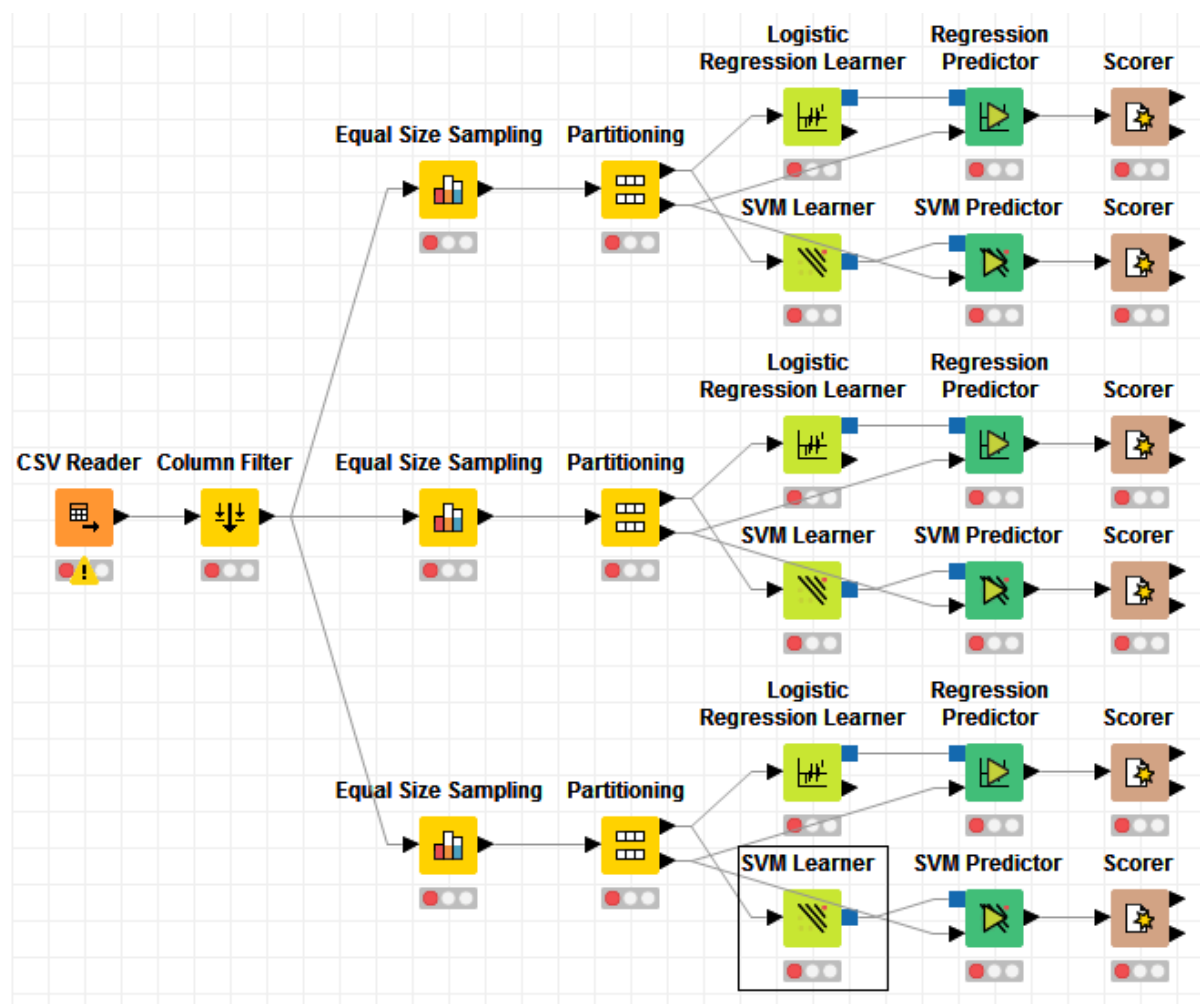


Figure 5-9. An example of a KNIME workflow illustrating the steps of analysis that were taken to compare SVM to LR. The CSV Reader node reads the data. The Column Filter node selects what variables will be included in the model fitting. The Equal Size Sampling node takes a sample of the data in which the two classes (solvate and nonsolvate) are equal in number. The Partitioning data node splits the data into a training set (10 % of the data) and a test set (90 % of the data). The Learner and Predictor nodes fits the model and performs the prediction, respectively. The Scorer node gives the % accuracy (percentage of correct predictions).

For each solvent's dataset, 3 types of models were fitted, these are models with the first variable (lowest p-value), a model with the first two variables and a model with the first three variables. With each of these, 3 equal size samples were used for training and testing the models. In other words, if we consider Figure 5-9 to show one workflow, 3 workflows were

conducted per solvent. For the logistic regression models, this was straightforward, as no parameters require adjustment. The performance of the logistic regression models with these three types of model was recorded.

5.3.4 Parameter adjustment of the RBF kernel

The last adjustment required before the analysis can take place is the parameters of the RBF kernel. The main two parameters to adjust are the sigma and the C parameter ⁷. The sigma parameter specifies the limit of the influence of one training point on the model (as given by the formula of the radial basis function), while the C parameter is a penalty term that specifies the softness of the margins of the SVM model.⁷ (See section 2.3.2.2). These parameters have no pre-determined value, where the optimal value is different for each training data. Moreover, their optimal value would change with the sample provided for the training dataset. For these reasons, these parameters are classically adjusted *via* grid search, although other faster methods are being suggested.⁸ Therefore; after the data was split into a training and a test set, a grid search was performed using an “optimization loop” in KNIME.

As has been shown in section 5.3.3, each solvent had 9 equal size samples taken from the parent data. Each of these equal size samples was partitioned into a training and a test set under a specific static seed (randomization seed). The sigma and C parameters were adjusted for each of these partitions, simply because the optimized parameters would not be ideal for other equal size samples or a differently partitioned sample. For this reason, the same dataset, equal size sampling and partitioning was used to fit the SVM and logistic regression model in each trial, as illustrated in Figure 5-10.

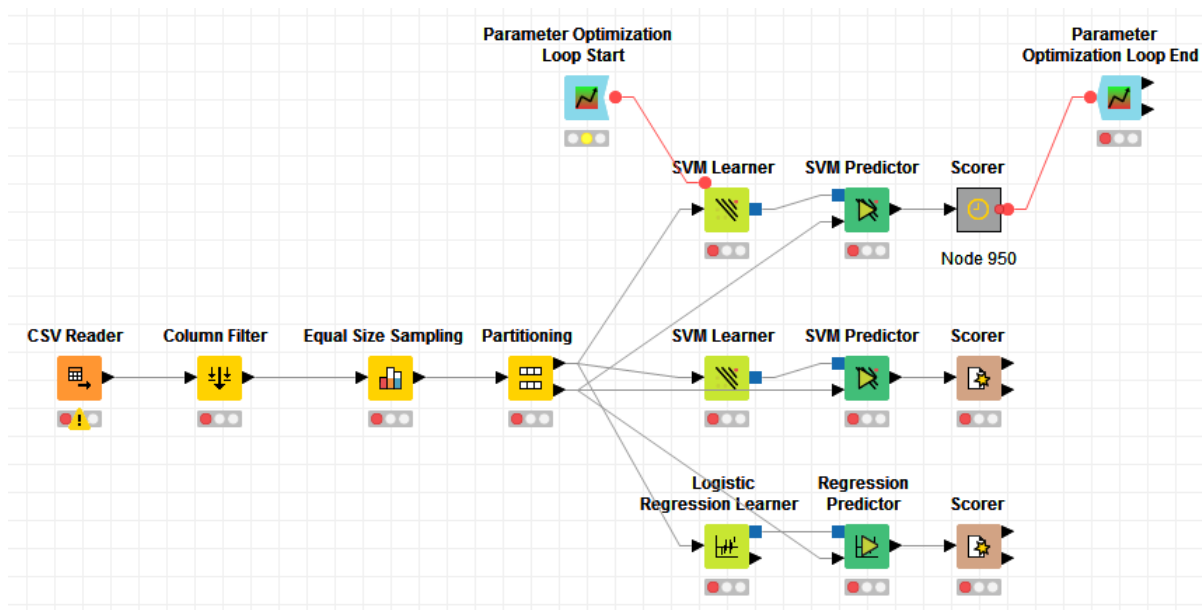


Figure 5-10. KNIME workflow for optimizing SVM parameters, testing an SVM model and testing a logistic regression model. This was done for 3 samples per variable which gives 9 trials for each of the solvents, resulting 45 similar workflows.

The workflow illustrated in Figure 5-10 represents the test that was done for 1 sample, where the results from the optimization loop (the first line in the work flow, which adjusted the sigma and the C parameter value) were used to fit the SVM model and compare it to the LR model. There are 5 solvents being tested, each of them had 3 types of models being tested, the 1 Variable, 2 Variable and 3 Variable models. In order to avoid the sampling error, each of these was tested 3 times, summing the total number of samples to 45.

5.3.5 SVM vs LR

The results of the 45 workflow are presented in bar plots in Figure 5-11 to Figure 5-15. Note that 45 workflow means 45 SVM and 45 LR model, resulting in 90 bars in total to compare.

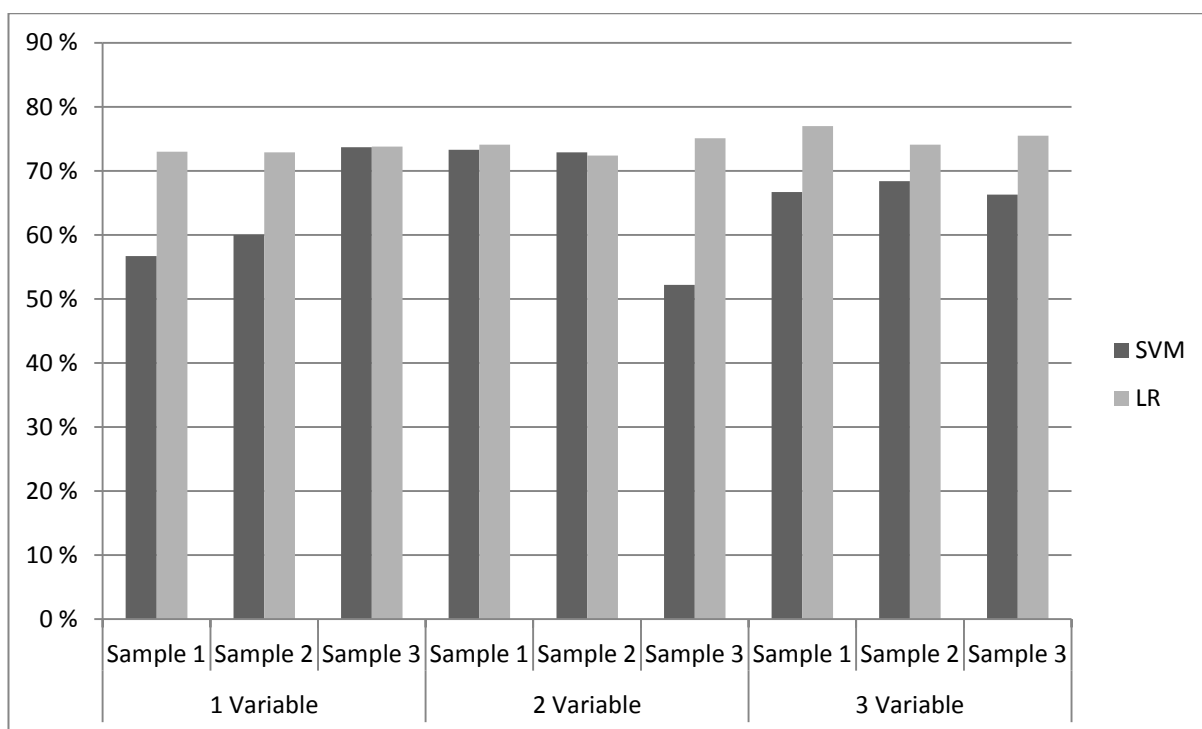


Figure 5-11. The percentage of correct predictions from the ethanol dataset made by the logistic regression and the support vector machine models that were fitted to the same samples. The test set is more than 1200 molecules.

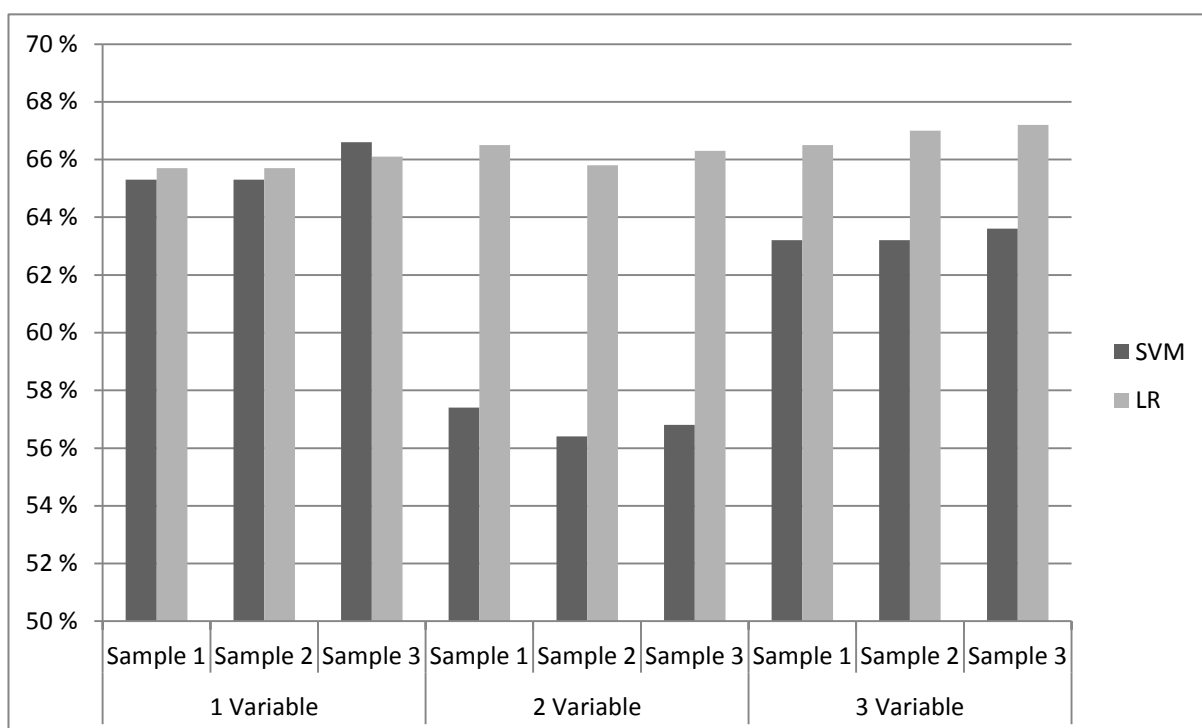


Figure 5-12. The percentage of correct predictions from the methanol dataset made by the logistic regression and the support vector machine models that were fitted to the same samples. The test set is more than 2700 molecules.

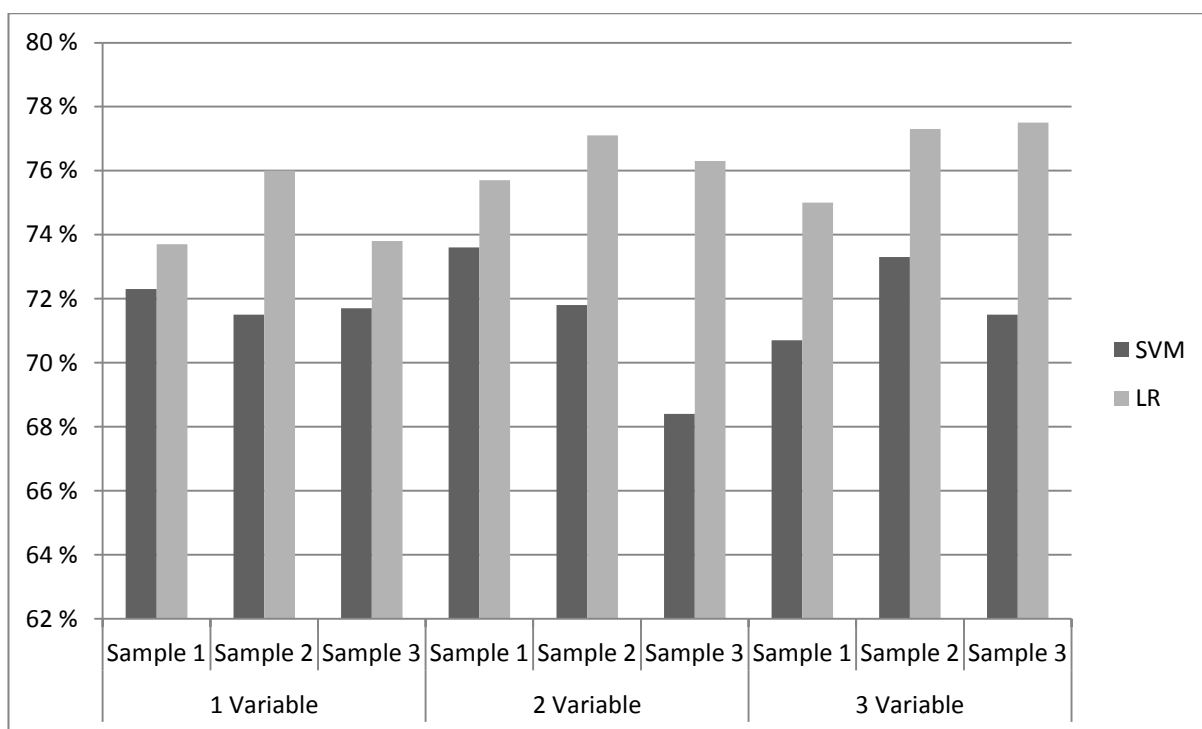


Figure 5-13. The percentage of correct predictions from the dichloromethane dataset made by the logistic regression and the support vector machine models that were fitted to the same samples. Test set is more than 2300 molecules.

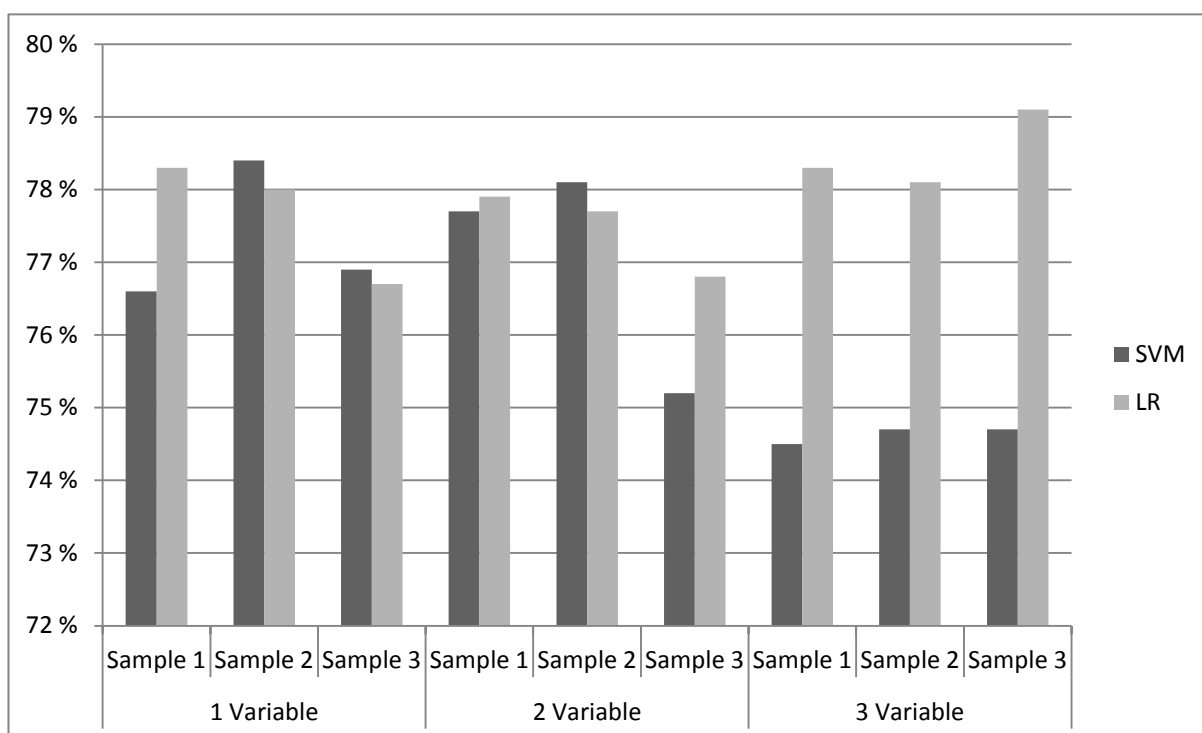


Figure 5-14. The percentage of correct predictions from the chloroform dataset made by the logistic regression and the support vector machine models that were fitted to the same samples. Test set is more than 2100 molecules.

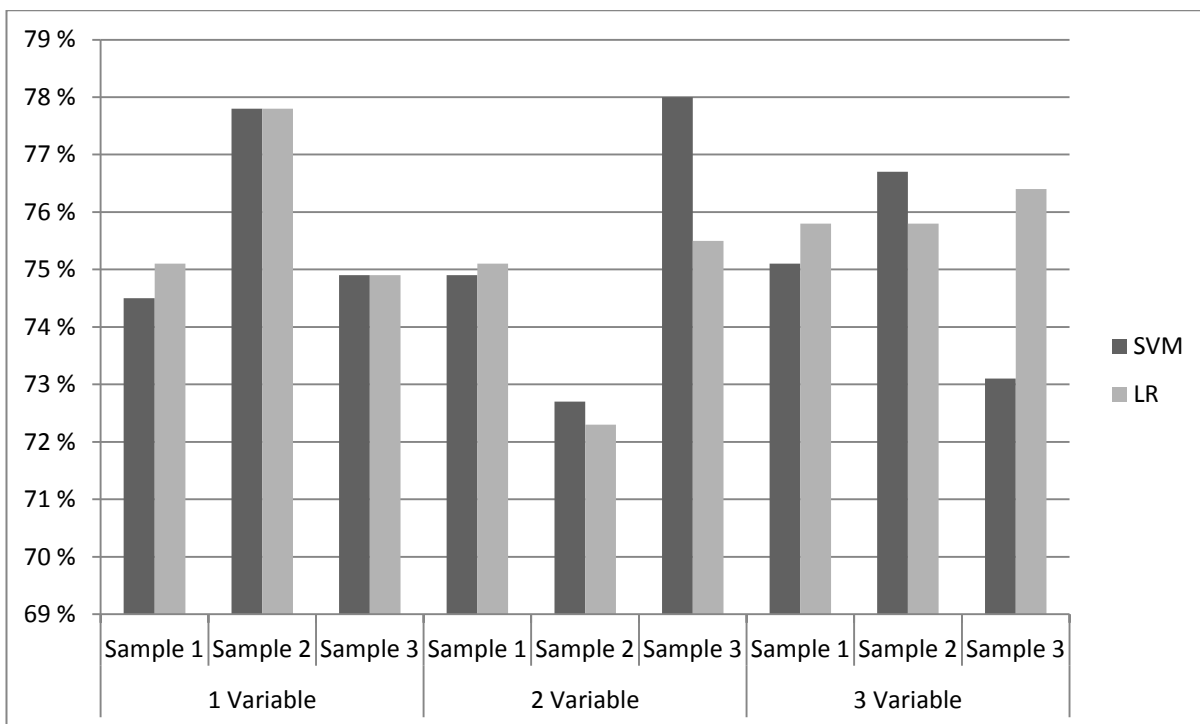


Figure 5-15. The percentage of correct predictions from the water dataset made by the logistic regression and the support vector machine models that were fitted to the same samples. Test set is more than 500 molecules.

Overall, LR has performed better than SVM in all solvents, where it showed higher prediction accuracy in most samples. SVM has topped LR in some cases, most notably the water 2 Variable model, Sample 3. Nevertheless, this is a small difference.

The fact that that LR mostly topped SVM as the number of dimension increased, signals that this problem is better classified using a linear classifier, or at the very least, it can be confidently said that there's no reason to use a more complex, non-linear method.

5.4 Principal components as logistic regression variables

Logistic regression turned out to be a good method of classification for the datasets under test, as non-linear SVM method showed no advantage over it. The question remains on what descriptors to include in this method to make predictions. Instead of looking for the descriptor that would give a good prediction, one idea was to start with what is already available. The PCA in section 5.1.2 has shown the ability to split the data into a solvate and a non-solvate forming groups (Figure 5-6). Although the main descriptors contributing to the formation of

these principal components were not found, the first three principal components were able to aggregate the data in what looked like two large clusters. This means these uncorrelated components could be used to make predictions if they were used as variables in a logistic regression function. The logistic regression function was applied in R, with the first three principal components being the variables used by the algorithm. The error rate associated with every model was estimated *via* cross-validation, where the cross-validation estimate of prediction error was used to decide the performance of the resulting models.^{9,10}

5.4.1 Models with one principal component

The PCA applied to the datasets here used equally sized samples for all solvent's datasets. For example in the ethanol dataset the total number of molecules approaches 5000, but only 1364 were used to perform the PCA to ensure that the principal components obtained explain the variability in the dataset of both the solvate and the non-solvate group. To estimate the sampling error, 10 equally sized samples were tested. Additionally, each model was cross-validated 10 times.

It is important to mention that all of these sampling procedures took place based on fixed seed values. This means the algorithm for randomly choosing these molecules is saved in the R script. This helps getting reproducible results when the script is run again.

The estimate through which the performance of the logistic regression model was measured was the Average Mean Squared Error of the 10-fold cross validation (MSE). Other estimates, such as the Akaike Information Criterion (AIC) and the Area Under the receiver operator curve ROC (AUC) were calculated. The analysis was started with the simplest possible form of the models; that is a logistic regression model that takes only the first principal component into account, denoted as model 1. The results of the models with the lowest MSE (best performance) and the models with the highest MSE (worst performance) among the 10 equal size samples for each solvent are shown in Table 5-3.

Table 5-3. Details of the logistic regression models taking into account the first principle component (model 1) in different solvents. The models with the lowest MSE (best performance) and the models with the highest MSE (worst performance) among the 10 equal size samples for each solvent are shown

Solvent	Performance	Number of datapoints	Mean squared error
Ethanol	Best	4893	0.187
Ethanol	Worst	4893	0.198
Methanol	Best	4364	0.213
Methanol	Worst	4364	0.217
Dichloromethane	Best	2759	0.156
Dichloromethane	Worst	2759	0.159
Chloroform	Best	2554	0.157
Chloroform	Worst	2554	0.160
Water	Best	4427	0.161
Water	Worst	4427	0.177

The models utilizing the first principal component only have shown a good separation ability of the data into the solvate and the non-solvate groups. This can also be noticed as the highest MSE value across all solvents was below 0.22. The inter-sample variation of the models is also little, where the largest difference in MSE between the 10 equal-size-samples was in ethanol, with a difference of 0.03. It is also notable that the AUC and the AIC always agree with the MSE value, indicating these terms can be used interchangeably for the estimation of the goodness of the fit. The AUC value can actually be visualized through a ROC curve. An example of the ROC curve of the ethanol model is shown in Figure 5-16.

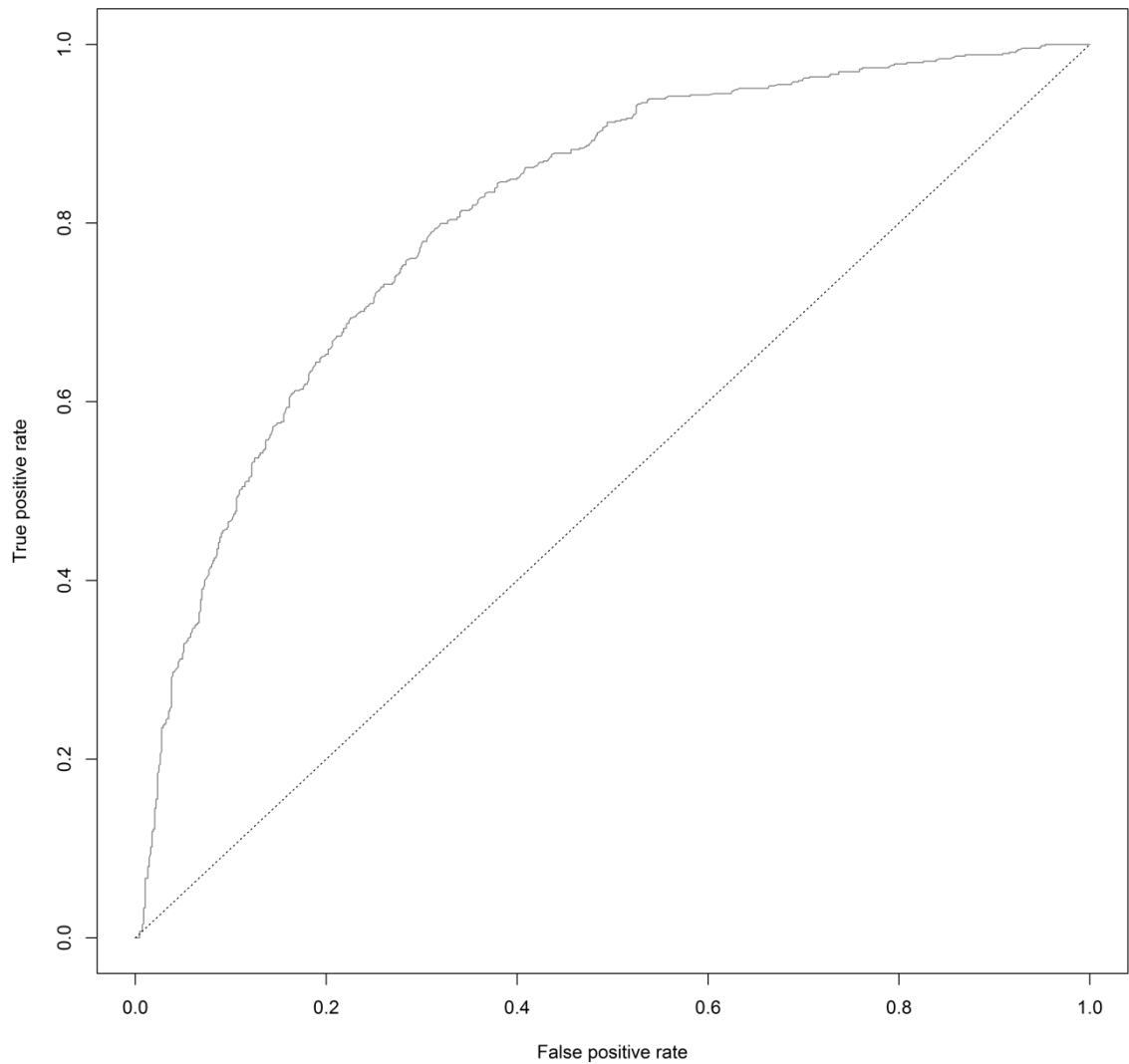


Figure 5-16. Receiver operative curve (ROC) of the ethanol model 1. The area under the curve is approx. 0.72. The dotted line, with the 45 ° angle represents the random guess (50 % correct prediction).

5.4.2 Models with two and three principal components

In order to see the improvement in the performance of the model upon the addition of next principal components, logistic regression models taking into account the first two and the first three variables (model3) were fitted. The results are given in Table 5-4 and Table 5-5, respectively.

Table 5-4. Details of the logistic regression models taking into account the first two principal components (Model 2) in different solvents. The models with the lowest MSE (best performance) and the models with the highest MSE (worst performance) among the 10 equal size samples for each solvent are shown

Solvent	Performance	Number of data points	Mean squared error
Ethanol	Best	4893	0.184
Ethanol	Worst	4893	0.194
Methanol	Best	4364	0.213
Methanol	Worst	4364	0.217
Dichloromethane	Best	2759	0.151
Dichloromethane	Worst	2759	0.154
Chloroform	Best	2554	0.153
Chloroform	Worst	2554	0.156
Water	Best	4427	0.161
Water	Worst	4427	0.178

Table 5-5. Details of the logistic regression models taking into account the first three principal components (Model 3) in different solvents. The models with the lowest MSE (best performance) and the models with the highest MSE (worst performance) among the 10 equal size samples for each solvent are shown

Solvent	Performance	Number of datapoints	Mean squared error
Ethanol	Best	4893	0.181
Ethanol	Worst	4893	0.192
Methanol	Best	4364	0.211
Methanol	Worst	4364	0.217
Dichloromethane	Best	2759	0.150
Dichloromethane	Worst	2759	0.153
Chloroform	Best	2554	0.153
Chloroform	Worst	2554	0.157
Water	Best	4427	0.161
Water	Worst	4427	0.177

Although the second and the third principal components accounted for part of the variability in the dataset (see section 5.1.2), they did not show much improvement to the predictive models. The addition of each principal component to the models has even shown an increase to the MSE value in some cases. For example looking at the water models, it can be seen that the MSE for the best model 1 has a lower MSE value than the best model 2 and model 3 all equal to 0.161. This shows that they do not add any advantage to the models based on model 1. This inability to show an improvement can be better illustrated through an overlay of the ROC curves, as shown in Figure 5-17.

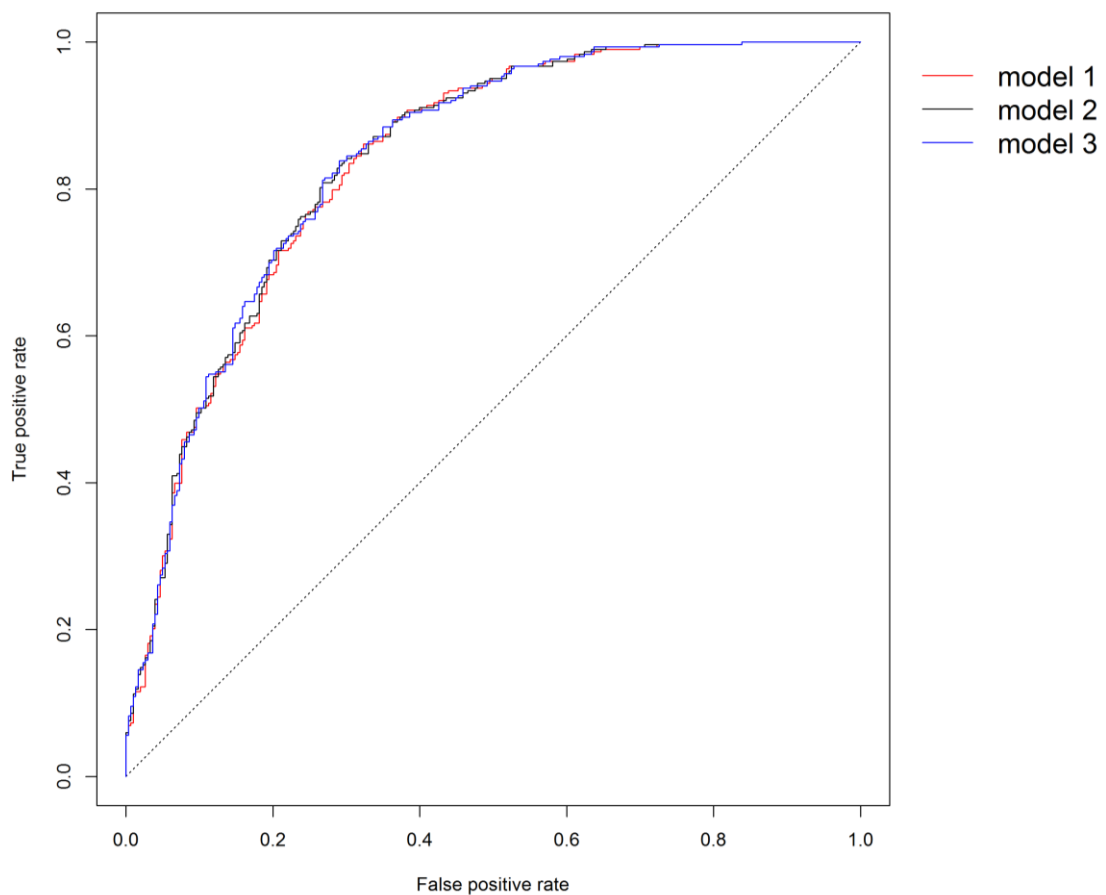


Figure 5-17. ROC curves of the water models taking into account 1, 2 and 3 principal components. The curves almost overlap, showing the insignificance of the addition of the second and the third principal components.

Bearing in mind that the MSE values shown in Table 5-3 to Table 5-5 are for best and worst models, which have the biggest difference among all 10 samples, the models prove to have a robust behaviour regardless of the subset that was taken from the original dataset.

5.5 Systematic variable selection using logistic regression

The first three PCs, each consisting of a combination of a few thousand descriptors, were able to give good classification ability with an MSE value between 0.150-0.217 in all solvents. Although this is a very good predictive ability, the principal components still do not give an idea on which of the descriptors actually contribute to the solvate formation. This is due to the low rotation values that were shown in Figure 5-8.

In order to see if there is a single, or a few descriptors that can classify the solvate and the non-solvate groups correctly, another approach was used. In this approach, only the descriptors that showed a significant difference between the solvate and the non-solvate groups in the Wilcoxon rank sum test were investigated. These descriptors were used in conjunction with the best machine learning algorithm found for this problem; that is the logistic regression.

5.5.1 Single-variable models

It is well known by now that about 3000 descriptors had significant effect on solvate formation. In order to select which among these give the logistic regression model that split the data best, a trial and error approach was used. Similar to the analysis conducted in section 5.4, the dataset of each solvent was split into 10 random equal size samples. For each of these samples, a number of models equal to the number of variables in each dataset were fitted. Each of these models was tested *via* a 10-fold cross validation and the MSE of the cross-validation was calculated. The variable that showed the lowest MSE was chosen as the best variable to describe the dataset. The performance of the models containing the best variable in each solvent is shown in Table 5-6.

Table 5-6. Single-variable models: details and performance. The models noted as “best” are the ones with the lowest MSE out of the 10 random samples. Similarly, the ones marked as “worst” have the highest MSE. This applies to Table 5-6, Table 5-9 and Table 5-10.

Solvent-model performance	Descriptors	Intercept	Descriptor 1 coefficient	No. of datapoints	MSE	AIC	AUC	% correct predictions	Cutoff point
Ethanol - best	SM5_H2	15.245	-2.067	1377	0.176	1486	0.816	73	0.485
Ethanol - worst	SM5_H2	13.458	-1.812	1376	0.191	1585	0.784	75	0.485
Methanol - best	SM3_Dt	6.486	-0.386	3034	0.204	3636	0.749	67	0.509
Methanol - worst	SM3_Dt	6.048	-0.358	3034	0.21	3709	0.732	68	0.508
Dichloromethane - best	SM3_H2	15.307	-3.219	2592	0.149	2419	0.866	78	0.515
Dichloromethane - worst	SM3_H2	14.798	-3.116	2592	0.152	2462	0.86	78	0.513
Chloroform - best	SM3_H2	14.319	-3.027	2384	0.152	2251	0.861	78	0.516
Chloroform -worst	SM3_H2	13.759	-2.915	2384	0.156	2307	0.854	78	0.514
Water- best	SpMaxA_Dt	3.771	-0.507	607	0.154	570	0.856	76	0.559
Water - worst	SpMaxA_Dt	3.016	-0.413	606	0.174	622	0.816	79	0.574

It is notable that all descriptors in these single-variable models were derived from the spectral moments of topological matrices. These are complex descriptors that represent the size and branching of molecules. Spectral moments are discussed in detail later (see section 5.6.2)

5.5.2 Two variable models

The single-variable models showed a surprisingly good ability to predict solvate formation. This lead to the anticipation that using a model that takes two variables into account rather than one would improve the predictive ability even further. The question was on how to select the best model that takes two variables into account? The ultimate answer to this question would be a trial-and-error approach. This means all possible two-variable models would be fitted, tested and compared. Logically, the descriptors that showed a significant difference between the solvate and the non-solvate groups in section 5.2.1 were used for this analysis. Fitting a model to every possible combination of length two for these ~3000 descriptors means that for each solvent, over 4 million models were fitted. In order to obtain unbiased models, ten equal size samples were used again here. This means that a total of more than 200 million models were fitted. A 10-fold cross validation was applied to each of these models, where the average MSE of these 10-folds was calculated. The combination of variables with the lowest MSE in each of the 10 subsets were recorded, as seen in Table 5-7.

Table 5-7. The 10 descriptor combinations with the lowest MSE in the dichloromethane data set – sample 1

Combination	MSE
SM3_H2 + Hy	0.14569
SM3_H2 + H.050	0.14588
EE_H2 + Hy	0.14596
SM3_H2 + nHDon	0.14597
SM4_H2 + Hy	0.14607
EE_H2 + H.050	0.14614
SM3_H2 + BLTD48	0.14619
SM3_H2 + MLOGP	0.14619
SM3_H2 + BLTA96	0.14619
SM3_H2 + BLTF96	0.14619

Doing this type of testing is CPU intensive. For this reason, these analyses were executed using the High Performance Computing Cluster at the University of East Anglia, where the analyses were parallelized between 160 dedicated cores with 64 GB of RAM. The script to perform the analyses was programmed in R, with the parallelization conducted using the “dopar” package¹¹. The time required for the analyses was between 3 and 5 days per dataset, depending on the number of data points. The models with the lowest errors among the 4 million were selected.

Due to the fact that 10 subsets were withdrawn from the complete dataset in each solvent, the best model obtained from each sample was different. For example, the best model might not necessarily utilize the same descriptors in each subset, as shown in Table 5-8.

Table 5-8. The descriptor combinations with the lowest MSE in each of the 10 subsets in the dichloromethane data

Sample No.	Combination	MSE
1	SM3_H2 + Hy	0.14569
2	SM3_H2 + Hy	0.14593
3	SM3_H2 + Hy	0.14585
4	SM3_H2 + Hy	0.14578
5	SM3_H2 + MLOGP	0.14565
6	SM3_H2 + Hy	0.14699
7	SM3_H2 + Hy	0.14694
8	SM3_H2 + Hy	0.14705
9	SM3_H2 + MLOGP	0.14699
10	SM3_H2 + MLOGP	0.14692

Although the 10 subsets of the Dichloromethane dataset resulted in different best models, only one combination should be chosen for the final models. Therefore the best combination was selected based on a voting system, where the pair of descriptors that gave the lowest MSE in most of the 10 samples were chosen. The resulting models are shown in Table 5-9.

Table 5-9. Two-variable models: details and performance

Solvent-model performance	Descriptors	Intercept	Descriptor 1 coefficient	Descriptor 2 coefficient	No. of datapoints	MSE	AIC	AUC	% correct predictions	Cutoff point
Ethanol - best	AVS_H2 + <i>nHDon</i>	16.781	-4.012	-0.941	1377	0.145	1253	0.875	80	0.485
Ethanol - worst	AVS_H2 + <i>nHDon</i>	14.628	-3.473	-0.822	1376	0.158	1345	0.852	81	0.485
Methanol - best	TRS + <i>nHDon</i>	2.893	-0.087	-0.636	3034	0.177	3222	0.816	74	0.509
Methanol - worst	TRS + <i>nHDon</i>	2.687	-0.079	-0.586	3034	0.184	3334	0.802	75	0.508
Dichloromethane - best	SM3_H2 + Hy	15.791	-3.381	-0.651	2592	0.145	2363	0.873	79	0.515
Dichloromethane - worst	SM3_H2 + Hy	15.33	-3.289	-0.661	2592	0.148	2404	0.868	79	0.513
Chloroform - best	SM3_H2 + <i>H.050</i>	15.103	-3.124	-0.381	2384	0.146	2184	0.871	79	0.516
Chloroform - worst	SM3_H2 + <i>H.050</i>	14.586	-3.022	-0.384	2384	0.15	2237	0.864	80	0.514
Water- best	π ID + <i>Mor05u</i>	5.386	-0.522	0.317	607	0.153	565	0.862	78	0.569
Water - worst	π ID + <i>Mor05u</i>	4.677	-0.449	0.294	607	0.164	603	0.839	79	0.563

Compared to the single-variable models, the addition of the second variable has improved the predictive ability of the models in some but not all solvents. The MSE values were reduced into the range of 0.145 to 0.184. The variable that was added to the single-variable models of ethanol and methanol was related to hydrogen bond donation. These variables have contributed to the MSE reduction of the models. In dichloromethane and chloroform, the new variable was related to hydrophilicity and heteroatoms connected to hydrogens, all of which are correlated to hydrogen bonding (correlation between these variables is above 0.95 in all datasets). In water on the other hand the variable was related to the 3D structure of the compound. The second variable in dichloromethane, chloroform and water, did not result in a large model improvement, an illustration of the MSE value of the single- and the two-variable descriptors is shown in Figure 5-18.

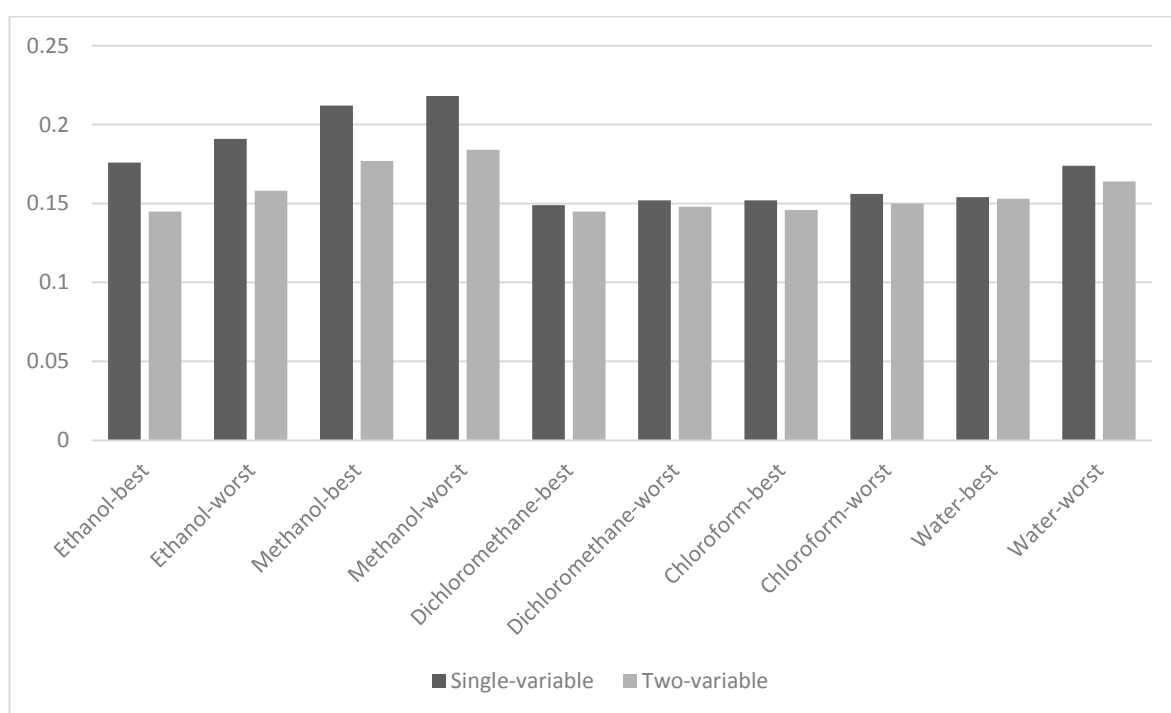


Figure 5-18. Reduction of the MSE by the addition of the second descriptor in each solvent.

5.5.3 Three-variable models

The two variable models have shown a large improvement over the single-variable models of ethanol and methanol in addition to a slight improvement in the predictive ability of the

dichloromethane, chloroform and water models. In order to improve the models even further, another descriptor was planned to be added.

The same exhaustive approach that was used to add the second descriptor (where all possible models of length two were tested), cannot be applied to the three-variable models. The number of possible models increases exponentially with the number of descriptors included. For example in the single variable models, around 3000 models were possible per solvent. Upon the addition of the second descriptor, the number of possible models increased to approximately 4.5 million per solvent. This required parallelization and the use of a cluster to perform this calculation. Using the same approach with three variables would increase the number of possible models to about 4.5 billion. Such a calculation would require a lot of computing power and it is not feasible to perform it within a reasonable timeframe. For this reason, the three-variable models were fitted based on the two-variable models *via* forward selection. The addition of a third variable to the two-variable models was performed using the “stats” package in R. This package offers the function “add1()” that fits models by adding one extra variable at a time to an existing model. Again here, since the datasets are not balanced in number between the solvate and the non-solvate groups, 10 equal size samples were used, resulting in a reasonable number (~30,000) of models per solvent. The “add1()” function evaluates the best model using the AIC value. But would our judgment be consistent if the AIC was used as a performance measure?

In order to answer this, Table 5-6 and Table 5-9 were checked again. In these tables, it can be seen that the AIC value always agreed with the MSE. Although these two estimates seem to generally agree on the performance of different models, a closer look is required to see how closely related they are. These two estimators are unit less and they use different approaches to be calculated. See section 2.3.3 for details on how the AUC and the MSE are calculated.

AIC values in the two-variable models fell between 565 and 3333 while the MSE values fell between 0.145 and 0.184. This means the former values is around 10000 times larger than the latter for the same model. In order to compare these two estimators, their values were normalized to their means, the comparison is shown in Figure 5-19.

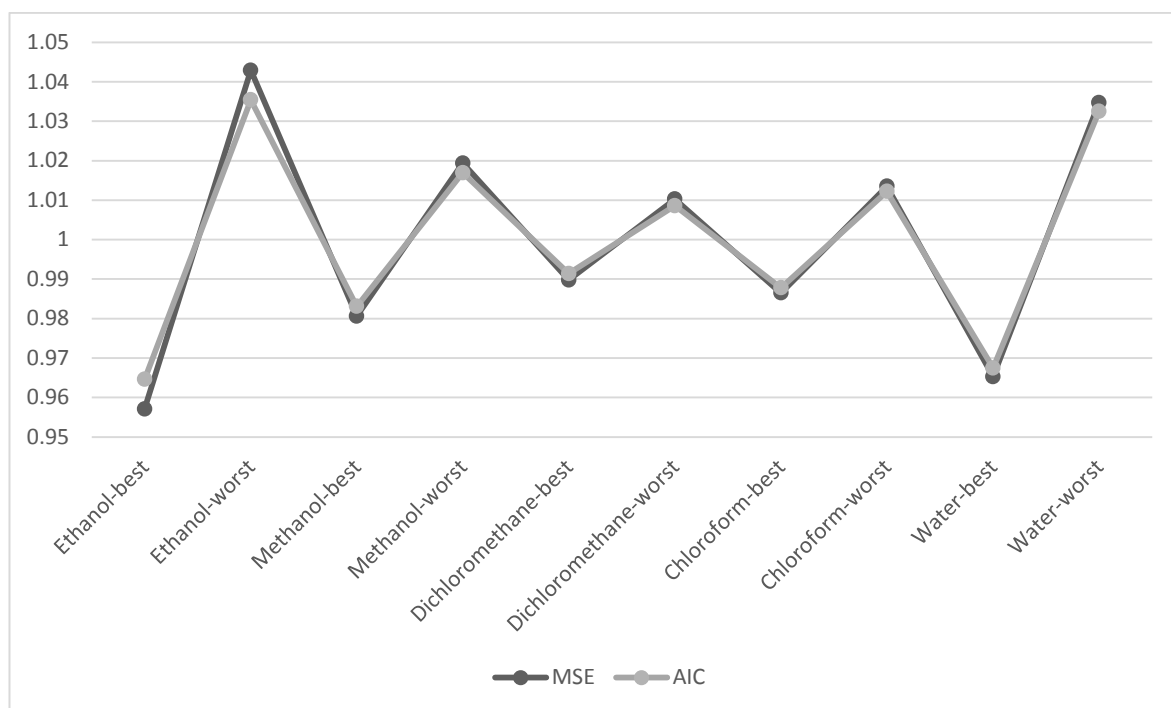


Figure 5-19. A plot of the normalized AIC and MSE value for the best and the worst two-variable models. The details of these models were shown in Table 5-9.

Figure 5-19 shows that the MSE and the AIC are very close estimates, supporting the idea that the AIC is as efficient as the MSE in these models. Therefore, the models with the lowest AIC values were selected after the “add1()” function was applied. The best and the worst model of these 10 samples in each solvent, along with the AIC, MSE and the AUC values are shown in Table 5-10.

Table 5-10. Three-variable models: details and performance

Solvent-model performance	Descriptors	Intercept	Descriptor 1 coefficient	Descriptor 2 coefficient	Descriptor 3 coefficient	No. of datapoints	MSE	AIC	AUC	% correct predictions	Cutoff point
Ethanol - best	AVS_H2+ nHDon+ Mor24u	16.745	-4.002	-0.941	0.024	1377	0.145	1255	0.875	80	0.485
Ethanol - worst	AVS_H2 + nHDon + Mor24u	14.557	-3.454	-0.821	0.047	1376	0.158	1347	0.852	81	0.485
Methanol - best	TRS + nHDon + SpMAD_AEA.ri.	2.709	-0.088	-0.636	0.127	3034	0.178	3224	0.815	74	0.509
Methanol - worst	TRS + nHDon + SpMAD_AEA.ri.	2.43	-0.08	-0.586	0.177	3034	0.184	3336	0.801	75	0.508
Dichloromethane - best	SM3_H2 + Hy + G2u	15.874	-3.453	-0.635	0.973	2592	0.142	2341	0.877	79	0.515
Dichloromethane - worst	SM3_H2 + Hy + G2u	15.238	-3.314	-0.621	0.855	2592	0.146	2388	0.871	79	0.514
Chloroform - best	SM3_H2 + H- O50 + nHBonds	15.498	-3.204	-0.563	0.414	2384	0.145	2168	0.874	79	0.516
Chloroform - worst	SM3_H2 + H- O50 + nHBonds	15.021	-3.11	-0.578	0.447	2384	0.149	2219	0.867	79	0.514
Water- best	π ID + Mor05u + F03.O.O.	5.39	-0.523	0.312	-0.017	607	0.153	567	0.862	78	0.569
Water - worst	π ID + Mor05u + F03.O.O.	4.378	-0.38	0.339	-0.015	606	0.166	598	0.837	78	0.574

While the AIC values of some solvents slightly decreased (such as dichloromethane), others have shown a slight increase in this value (e.g. ethanol). The AIC values of the new models were generally similar to those of the two-variable models. This leads to the conclusion that the addition of the third variable increased the complexity of the model without giving extra information that is good enough to outperform the complexity. The MSE values were similarly affected. As the MSE was used to evaluate previous models, a comparison of the MSE values in the two- and the three-variable models is shown in Figure 5-20.

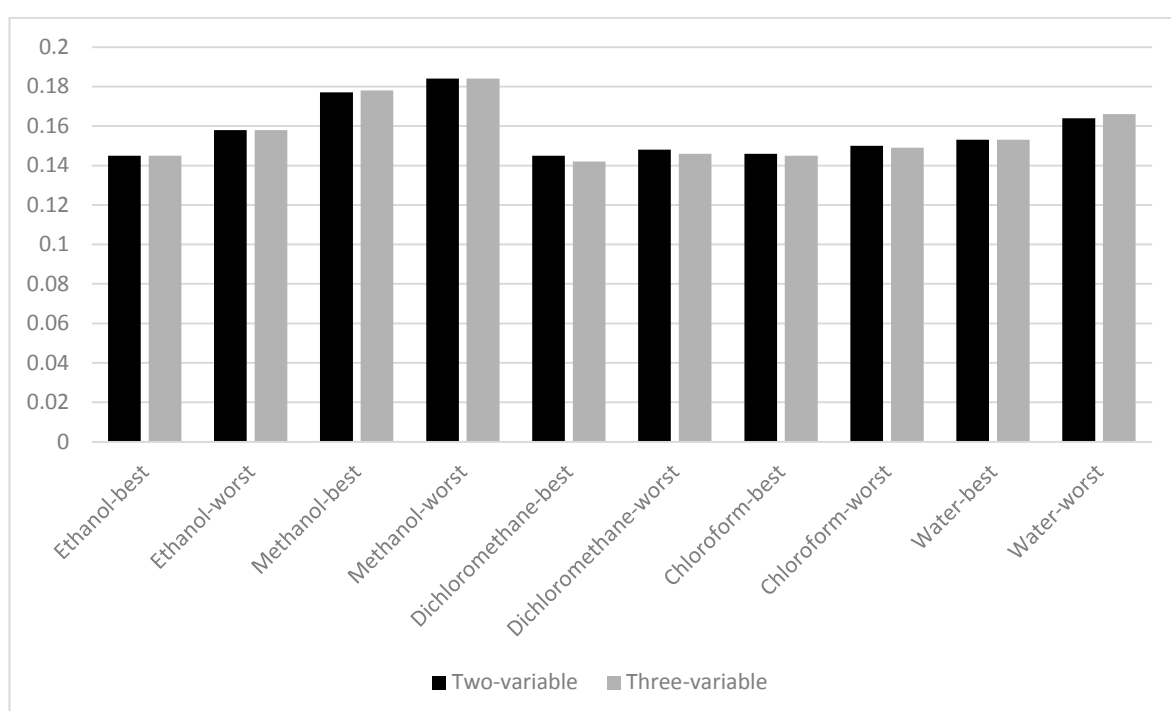


Figure 5-20. The change in MSE between the two-variable and the three-variable models. For each solvent, the models with the lowest and the highest MSE of the 10 equally sized samples is shown.

The addition of more descriptors to the current models does not seem to improve their predictive ability any further. For this reason, the two-variable models are going to be used.

5.6 A closer look on the two-variable models

After all the investigations were conducted in section 5.4., the two-variable models turned out to give the most reasonable predictions for solvate formation, both in terms of complexity and

performance. Therefore, these models are going to be used to make predictions. In this section, the two-variable models are going to be presented, discussed and analysed.

5.6.1 The models

The general formula for the predictive models is given in Equation (5-1).

$$Probability = \frac{1}{1 + e^{-(Intercept + Coefficient1 * Descriptor1 + Coefficient2 * Descriptor2)}} \quad (5-1)$$

Each model contains two types of parameters, the constants, which only change when predicting a different solvent's behaviour and the descriptors, whose values change depending on the molecule to be predicted. For the constants part, the intercept showed nonnegative values in all models. On the other hand, the coefficients of all descriptors had negative values except for one descriptor in the water model. It is important to keep in mind that the descriptor values themselves can be positive or negative. Each descriptor is multiplied by its coefficient. In logistic regression, if the product of the descriptor and its coefficient gives a negative term, it pushes the probability value towards zero (0), while a positive one pushes the prediction towards one (1). In these models, solvate formation is more likely when the probability is closer to zero, while non-solvate formation is more likely when the prediction is closer to one.

Although the descriptors to include in the model of each solvent are known at this point, the coefficients of these models would slightly differ depending on the sample used for training the algorithm. In order to obtain models that are fair, 10 models were obtained for 10 random subsets per solvent (these are the same 10 subsets used earlier in cross-validation). The

models were then averaged in terms of the intercept, descriptor 1 coefficient and descriptor 2 coefficient. These models are given in Equations (5-2) to (5-6).

$$p_{ethanol} = \frac{1}{1 + e^{-(15.939 - 3.817AVS_H2 - 0.861nHDon)}} \quad (5-2)$$

$$p_{methanol} = \frac{1}{1 + e^{-(2.8 - 0.085TRS - 0.612nHDon)}} \quad (5-3)$$

$$p_{dichloromethane} = \frac{1}{1 + e^{-(15.459 - 3.314SM3_H2 - 0.664Hy)}} \quad (5-4)$$

$$p_{chloroform} = \frac{1}{1 + e^{-(14.744 - 3.05SM3_H2 - 0.384H050)}} \quad (5-5)$$

$$p_{water} = \frac{1}{1 + e^{-(4.672 - 0.424\pi ID + 0.327Mor05u)}} \quad (5-6)$$

where $p_{ethanol}$, $p_{methanol}$, $p_{dichloromethane}$, $p_{chloroform}$ and p_{water} are the probability of a molecule to stay in the non-solvated form when crystallised from the corresponding solvent. The result is a value between 0 and 1, where 0 represents a solvate and 1 represents a non-solvate.

Since these models, along with the coefficients they contain, were averaged over multiple datasets, a reasonable question would be how much confidence is associated with each of these coefficients. In other words, what was the deviation of each sample from the mean value of the coefficient that is shown in the Equations (5-2) to (5-6)? The confidence of the intercept, the first and the second coefficient in each model can be represented by finding the standard deviation of the 10 models, fitted to the subsets of the complete data. Comparing the absolute standard deviations won't give an intuitive value. This is because some coefficients have large and others have small values for standard deviations. In order to make the values comparable, the relative standard deviation (also known as the coefficient of variation) can be used instead, that is the standard deviation divided by the mean.¹² The relative standard deviations are shown in Table 5-11.

Table 5-11. Relative standard deviation of the intercept, first coefficient and second coefficient over 10 models in each solvent

Solvent	Intercept	Coefficient 1	Coefficient 2
Ethanol	0.04	0.05	0.05
Methanol	0.03	0.03	0.04
Dichloromethane	0.01	0.01	0.04
Chloroform	0.01	0.01	0.02
Water	0.08	0.13	0.08

As it can be noticed, the deviation among the 10 samples was minimal. The largest deviation that can be seen is the first coefficient of the water model. This is expected, as the water had the most imbalanced sample, leading to almost completely different subsets that were used to fit the model, when equal size sampling was used.

5.6.2 The meaning of the descriptors

In this section, the descriptors that are utilized in each of the two-variable models are going to be discussed in details. For the ethanol model, the predictive model utilizes the AVS_H2 and nHDon descriptors. AVS_H2 is a descriptor that is determined based on the reciprocal squared topological distance matrix, and it is calculated by taking the natural logarithm of the average of the sum of the entries in each row of the matrix. The details of calculating this descriptor are provided in section 3.2.1.2. The AVS_H2 descriptor value is linked to molecular size and branching in a molecular graph. A molecule that is large and highly branched is expected to have a large AVS_H2 value. The second descriptor in the best ethanol two-variable model was a simple count descriptor, nHDon. This descriptor simply counts number of hydrogen bond donors *via* the molecular graph. These are hydrogen atoms that are bound to a nitrogen or an oxygen atom, according to the Dragon software documentation.

In the methanol two-variable model, the first descriptor was TRS (Total Ring Size). This is the total number of atoms in each independent ring in the molecule (e.g. TRS value of benzene is 6 and of naphthalene is 12). The fact of having this descriptor as part of the best methanol model could indicate the role of the hydrophobic ring interactions in the stabilization of solvate crystals. The second descriptor in the methanol model was the same as the second descriptor in the ethanol model (nHDon: the number of hydrogen bond donors).

The dichloromethane model had the SM3_H2 and the Hy descriptors. The former refers to the third order spectral moment of the reciprocal squared distance matrix (H2).¹³ The third spectral moment is calculated as the trace of the third power of the matrix.^{14 844–849.} Since this descriptor is calculated from the H2 matrix, it is closely related to the AVS_H2 descriptor observed in the ethanol model, where it also incorporates information about the size and branching of molecules. In fact, SM3_H2 is obtained by the logarithmic transformation [$x' = \ln(1+x)$] of the spectral moment. This is due to the exponential increase in its value with

the molecular size. The second descriptor in the dichloromethane model was H_y ; the hydrophilic factor. This factor is calculated using the formula presented in Equation (5-7):

$$H_y = \frac{(1+N_{Hy}) \cdot \log_2(1+N_{Hy}) + nC \cdot \left(\frac{1}{nSK} \log_2 \frac{1}{nSK} \right) + \sqrt{\frac{N_{Hy}}{nSK^2}}}{\log_2(1+nSK)} \quad (5-7)$$

N_{Hy} is the number of hydroxyl, amine or thiol groups, nC is the number of carbon atoms and nSK is the number of non-hydrogen atoms.¹⁵

The chloroform model shares the same first descriptor with the dichloromethane model; that is SM3_H2. This descriptor, combined with H-050 resulted in the chloroform model with the lowest MSE. The descriptor H-050 is calculated by counting the number of hydrogen atoms attached to a heteroatom.^{16, 17} This descriptor is highly correlated to H_y ($r > 0.95$). This suggests a proximity in behaviour between dichloromethane and chloroform in terms of solvate formation, where they have the first descriptor in common and have a near-identical second descriptor.

The best water two-variable model, included the πID and the Mor05u descriptors. The former is based on the conventional bond order ID number.¹⁸ It is calculated using the formula in Equation (5-8):

$$\pi ID = \ln(1 + nSK + \sum_p w_p) \quad (5-8)$$

were nSK is the number of non-hydrogen atoms and w_p is the weight of molecular path p . The index p runs over all bond paths in the hydrogen-depleted molecular graph, where the length of the path ranges from 1 bond to the longest path possible. Each path p is weighted by conventional bond orders of the bonds in this path, resulting in w_p ; the weight of molecular path. The conventional bond order of single bonds is 1, 1.5, 2 and 3 for single bonds, aromatic bonds, double bonds and triple bonds, respectively. By incorporating this information, the πID descriptor value doesn't only incorporate information regarding the complexity (size and branching) of a molecule. It also indicates the rigidity of a molecule. Note that the descriptor was subject to logarithmic transformation [see Equation (5-8)]. The second descriptor in the water model was a 3D-MoRSE (3D-Molecule Representation of Structures based on Electron diffraction) descriptor; that is Mor05u. This family of descriptors are calculated from the atomic 3D coordinates obtained by a molecular transform that is similar to the electron diffraction formulae,¹⁹ the formula to calculate the Mor05u descriptor is presented in Equation (5-9):

$$Mor05u = \sum_{i=1}^{nAT-1} \sum_{j=i+1}^{nAT} \frac{\sin(5r_{ij})}{5r_{ij}} \quad (5-9)$$

where r_{ij} is the topological distance between atoms i and j in the molecule and nAT is the total number of atoms in a molecule.

The 3D-MoRSE descriptors require the previous knowledge of the 3D coordinates of atoms in a molecule before conducting the descriptor calculation. This is not always applicable, especially that this model aims at predicting solvate formation at early development stages. For this reason, the Mor05u in the water model can be replaced with a 2D descriptor, providing the ability to predict solvate formation using the molecular graph only. A highly

correlated, easy-to-calculate descriptor can be used instead. Fortunately, two simple descriptors showed high correlation ($r = 0.94$) with Mor05u; these are the number of hydrogen atoms (nH) and the number of atoms of molecule (nAT). A water model based on the πID and nH descriptors gives an average MSE of 0.161 compared to an average MSE of 0.159 of the original hydrate model. The alternative water model is given in Equation (5-10).

$$p_{water-alternative} = \frac{1}{1 + e^{-(4.756 - 0.434\pi ID - 0.089nH)}} \quad (5-10)$$

With exception to water, all two-variable models consisted one descriptor that related to the size and branching of a molecule and another one that is related the heteroatoms in the molecules. The water model on the other hand has possessed one variable related to size and branching in addition to rigidity, while the second variable focused on the count of hydrogen atoms in a molecule.

5.6.3 The descriptor values and their coefficients

The models of the different solvents include descriptors that are identical or highly correlated, as has been shown in section 5.5.1, Equations (5-2) to (5-6). In the first instance, this could imply that the models are similar. In fact, it shows that the models are conceptually, but not numerically similar. The value of parameters in the logistic function (i.e. intercepts and coefficients) vary widely between the different models. This difference can be demonstrated using the ethanol and methanol models, where the relative importance of a shared descriptor ($nHDon$) is roughly 1.5 times higher in the ethanol model than it is in methanol model.

Each descriptor that is included in the logistic regression equation has a different impact on the overall result. The influence of each descriptor on the total model can be estimated by

looking at two properties: the coefficient of each descriptor and the value of the descriptor itself. Looking at the general logistic regression formula [Equation (5-1)], the value of each descriptor is multiplied by the coefficient found by the model. For example: in the ethanol model, for a given molecule, the value of the AVS_H2 descriptor is multiplied by -3.817. The mean value of this descriptor across the dataset is 3.728. Their multiplication gives a value of -14.23. The product of the descriptor value and its coefficient also has to be compared to the product of the other descriptor and its coefficient. The coefficient of the number of hydrogen bond donors (nHDon) descriptor is -0.861 and the mean value of this descriptor across the dataset is 1.113. The effect of the mean value multiplied by the coefficient gives a value of -0.958. This means that for an “average” molecule, the value of the AVS_H2 descriptor has an effect on the model around 15 times the value of the number of hydrogen bond donors (nHDon). This clearly shows that the importance of the size, complexity and branching exceeds the importance of the hydrogen bonding ability of a molecule in forming ethanol solvates. The negative sign in both descriptors indicates that larger values of the descriptors push the prediction value closer to zero i.e. towards solvate formation.

The value of a coefficient alone in the model is not enough to imply the importance of a descriptor as the value of the descriptor itself might be small or large. For example in the methanol model, the coefficients of the TRS and the nHDon descriptors are -0.084 and -0.612, respectively. In the first instance one might think that the hydrogen bonding is more important than the size for determining the methanol solvate formation ability. However, the mean value of the methanol model descriptors was 22.328 for the TRS descriptor and 1.497 for the nHDon. The product of the multiplication of the mean values with the coefficients is -1.876 for TRS and -0.916 for nHDon. This leads to the conclusion that the size and branching of a molecule is still the most important factor in comparison to the hydrogen bonding in methanol. Care must be taken not to take the coefficient or the descriptor value as an indicator of the final probability as it might cause misinterpretation of the results.

5.6.4 Visual representation

Since each solvent's model equation contains two variables, it is possible to represent these models in a 2D plot. An illustration of the dichloromethane model, its decision boundary and a sample's prediction is shown in Figure 5-21.

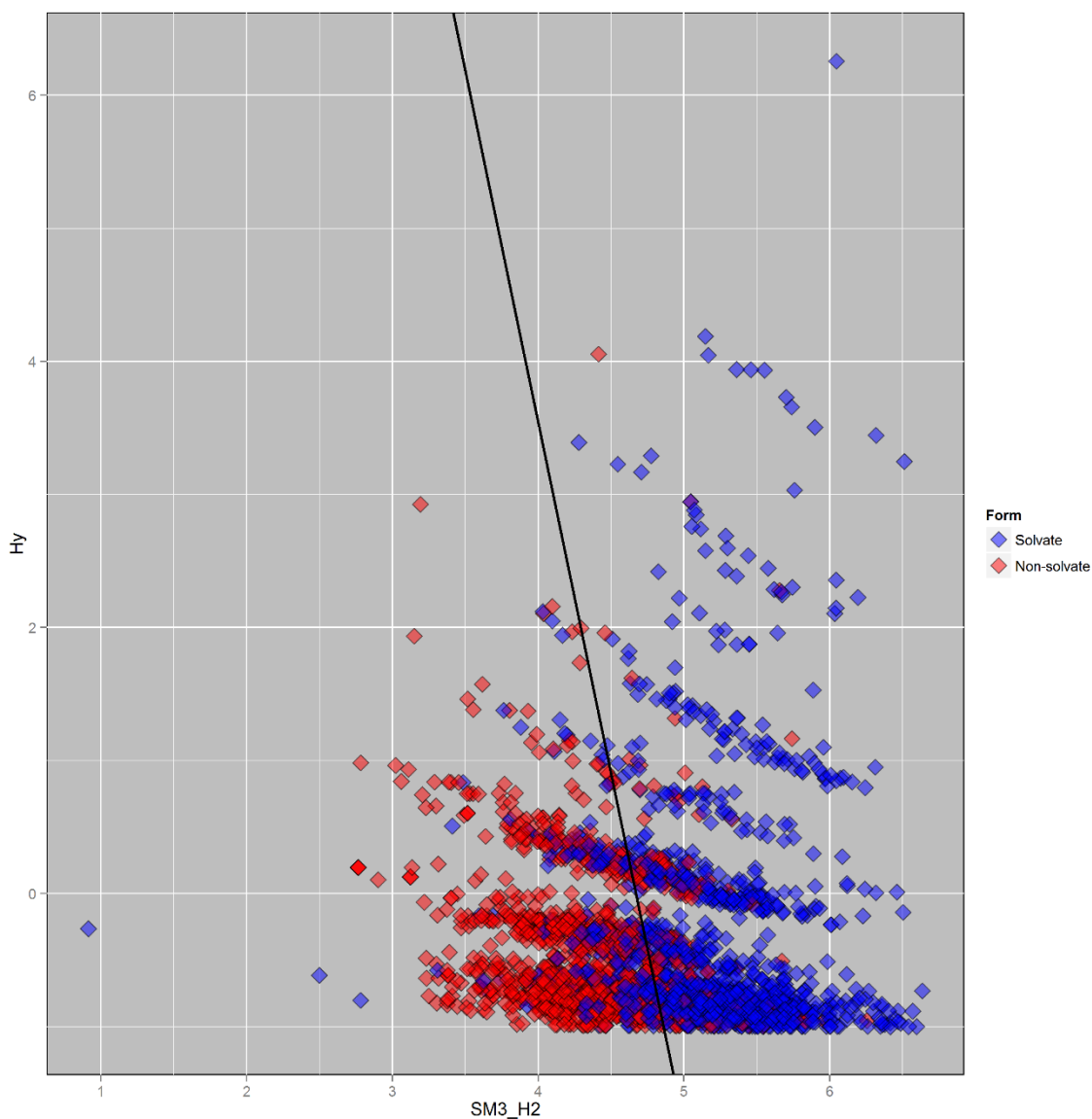


Figure 5-21. A plot of 2600 datapoints from the dichloromethane dataset in terms of the 2 descriptors that give the best linear separation, these are the SM3_H2 and Hy. The black line represents the decision boundary upon which the outcome is predicted. Colours = experimental outcome.

The x axis in Figure 5-21 represents the SM3_H2, a descriptor explaining the size and branching of a molecule. The increase in number of solvate with increased SM3_H2 suggests that a large,

branched molecule is more likely to form a solvate, probably because it would be difficult for such a molecule to optimally fill the three-dimensional space. It is also possible that poor packing of molecules in the crystal allows the solvent molecules to diffuse through the structure and form a solvate. This idea could be illustrated by a histogram of the SM3_H2 descriptor in the solvate and the non-solvate groups, as shown in Figure 5-22.

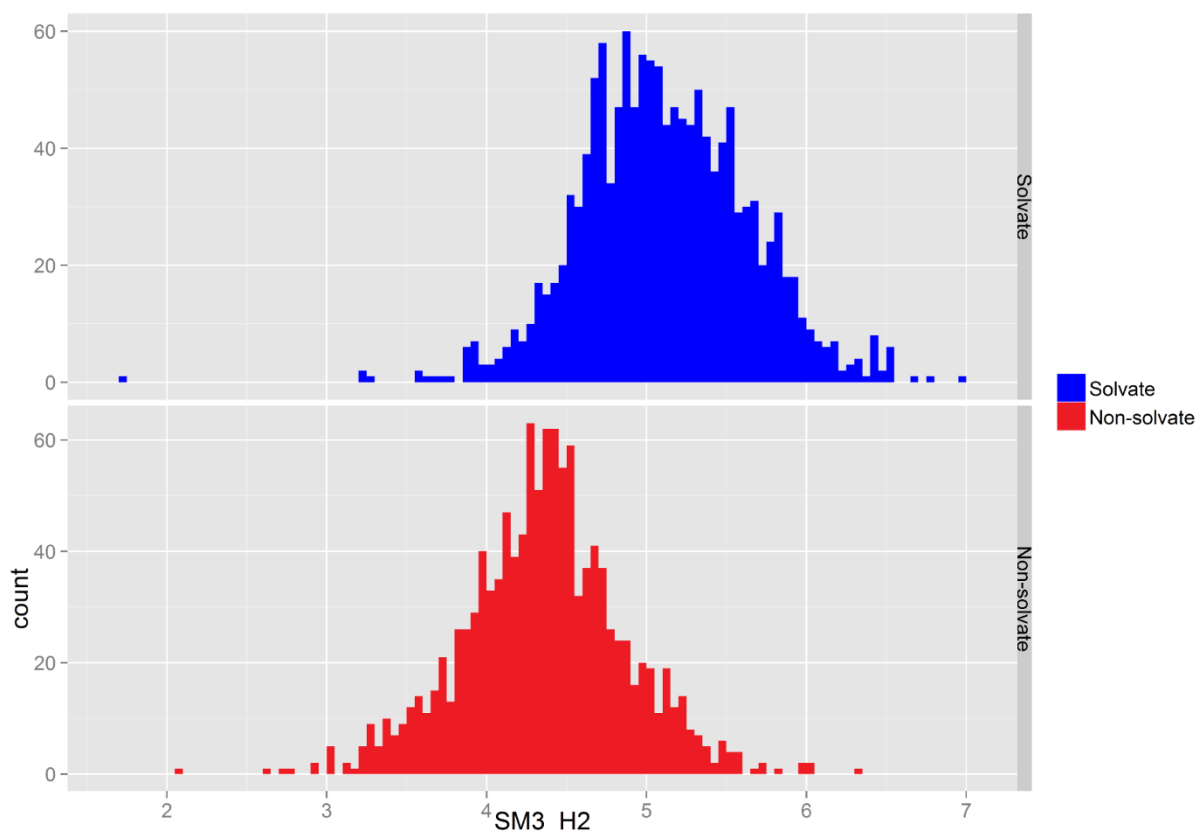


Figure 5-22. Histograms of the SM3_H2 descriptor distribution from the chloroform data.

The y axis in Figure 5-21 represents the number of hydrogen bond donors. The increase of number of solvents as this value increases suggests that hydrogen bond donors play a role in stabilizing the solvent molecules in the crystal voids. This can be supported by the fact that the hydrogen bonding related descriptors show a negative sign in the models, therefore contributing to solvate formation. The role of hydrogen bonding in solvate formation has been previously recognized.^{20, 21} The improvement in the current findings is the quantification of the

relative importance of size, branching and hydrogen bonding. By knowing this information, it would be possible to predict the ability of molecules to form solvate based on the molecular structure alone.

In the alcohol containing solvents, the hydrogen bonding related descriptors have shown a positive effect on solvate formation. Surprisingly, introduction of the hydrogen bond-related descriptors did not improve the predictive ability of the models for hydrate formation. The effect of the addition of a hydrogen bond related descriptor to the model is shown in Figure 5-23.

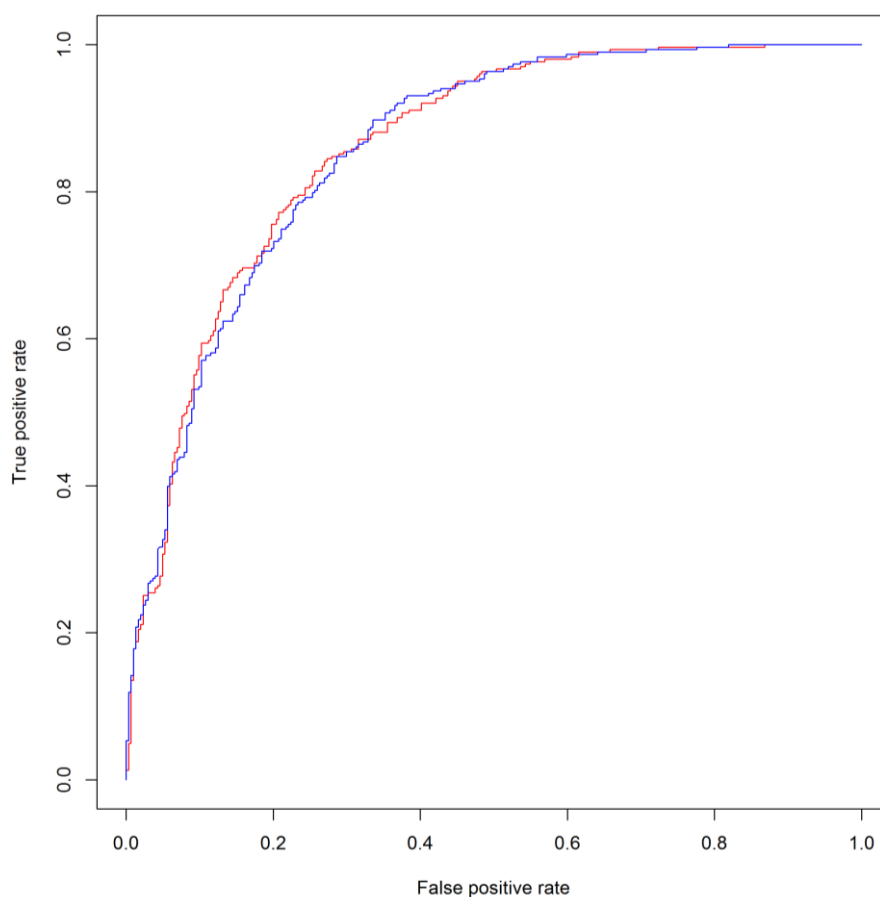


Figure 5-23. An illustration of the change in performance between the alternative water model fitted using piID and nH (red) vs the water model fitted using piID, nH and nHDon (blue). The steps in the curve (not smooth) are due to the small number of datapoints in the water sample (number of datapoints is 606).

In the first instance, it might seem that hydrogen bonding is not important in hydrate formation. The role of hydrogen bonding in hydrate formation is a fact that cannot be denied. Positive correlations between hydrogen bonding and hydrate formation has been reported.²² In order to demonstrate the importance of hydrogen bonding, a logistic regression model was fitted one variable only, that is the hydrogen bond acceptors (nHAcc). This model resulted in an average MSE of 0.237. This proves that hydrogen bonding plays an important role in hydrate formation, but it is not the most important discriminating factor according to this dataset. Note that all the datapoints provided for the water model were for crystals that were successfully grown from aqueous solutions suggests that even the non-hydrate formers among them are relatively hydrophilic.

The solvents that possess similar functional groups had their best models describing the same, or closely related structural features, although the datasets for all these solvents were completely different. For example, ethanol and methanol are both hydroxyl-containing solvents and the hydrogen bonding is essential in their solvate formation, as can be seen by the second descriptor in both. This is logical since both solvents are structurally related and can involve in similar interactions. Another example is the models of the dichloromethane and chloroform where the first descriptor was exactly the same (SM3_H2), chosen from amongst about 5000 descriptors. Such similar results agree with the expected outcome of similar behaviour of structurally related solvents.

5.6.5 Cut-off point determination

The cut-off point is the numerical probability value which splits the prediction outcomes into solvates and non-solvates. In principle, any value between 0 and 1 can be chosen to be the cut-off value. In order to get an unbiased decision, the convention is to choose a value that maximizes the true positive and the true negative predictions.²³ An illustration of the specificity and the sensitivity curves of the two-variable models is given in Figure 5-24.

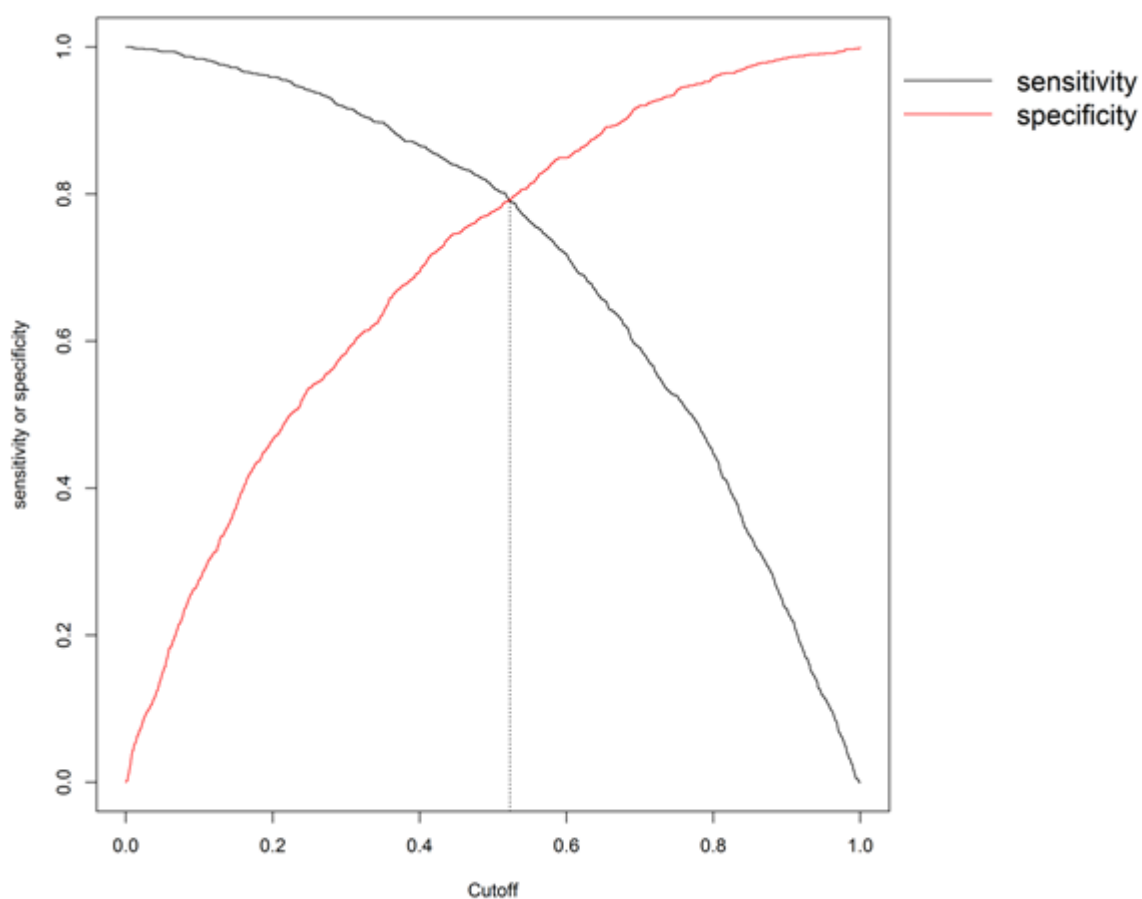


Figure 5-24. The sensitivity and specificity curves from the dichloromethane data, sample 1.

It is not too surprising that the optimal value found by crossing these two curves is close to 0.5, due to the fact that equal size samples were used. The value shown in Figure 5-24 is for one of the 10 samples taken per solvent. To get the best estimation possible for this point, the average cut-off point obtained for the 10 samples for each solvent was found. Table 5-12 shows the average cut-off point for each solvent.

Table 5-12. Average cut-off point in each solvent's dataset

Solvent	Average Cut-off point
Ethanol	0.485
Methanol	0.485
Dichloromethane	0.513
Chloroform	0.514
Water	0.564

The average cut-off point was very close to 0.5. The small deviation from 0.5 could be due to the sampling error, since this is an average over 10 samples. For this reason, 0.5 was chosen to be the cut-off point.

5.6.6 Residuals

One of the most important diagnostics of the model performance is the residuals. These are the difference between the actual value and the predicted value.²⁴ The importance of this diagnostic comes from the fact that the presence of a pattern among the residuals signals that the model is modifiable. On the other hand, if the residuals plot was fuzzy and showed no trend, then that means the model doesn't seem to be modifiable despite having prediction errors. The distribution around the model should also be symmetrical for a healthy residual plot, i.e. the points should be at comparable distances from the line at zero. The residual plot of the two-variable ethanol model is shown in Figure 5-25.

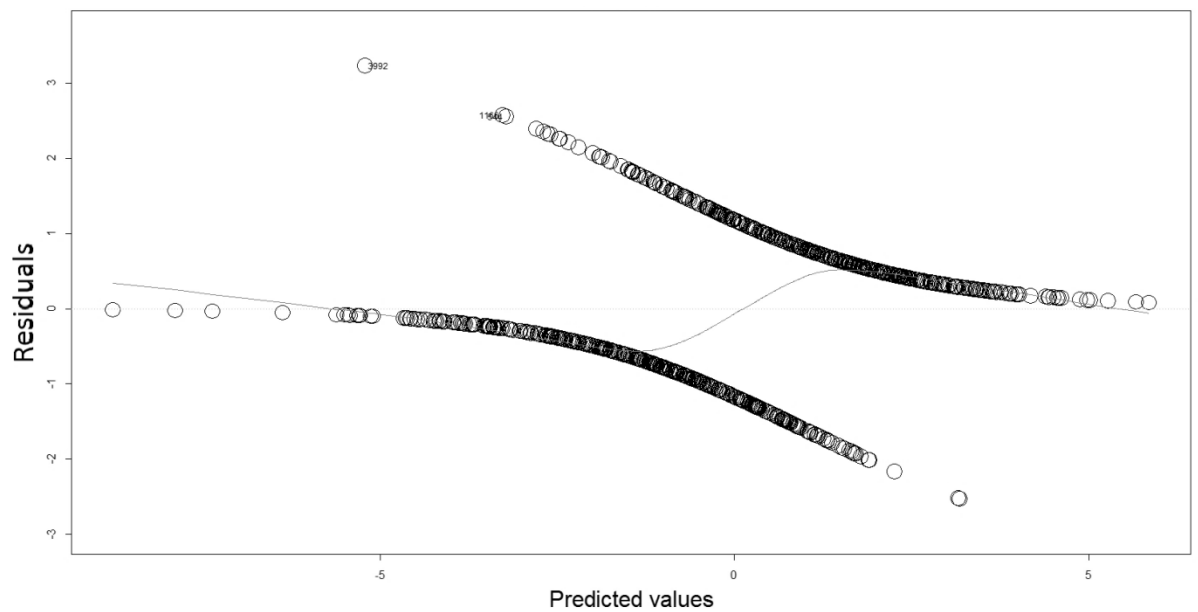
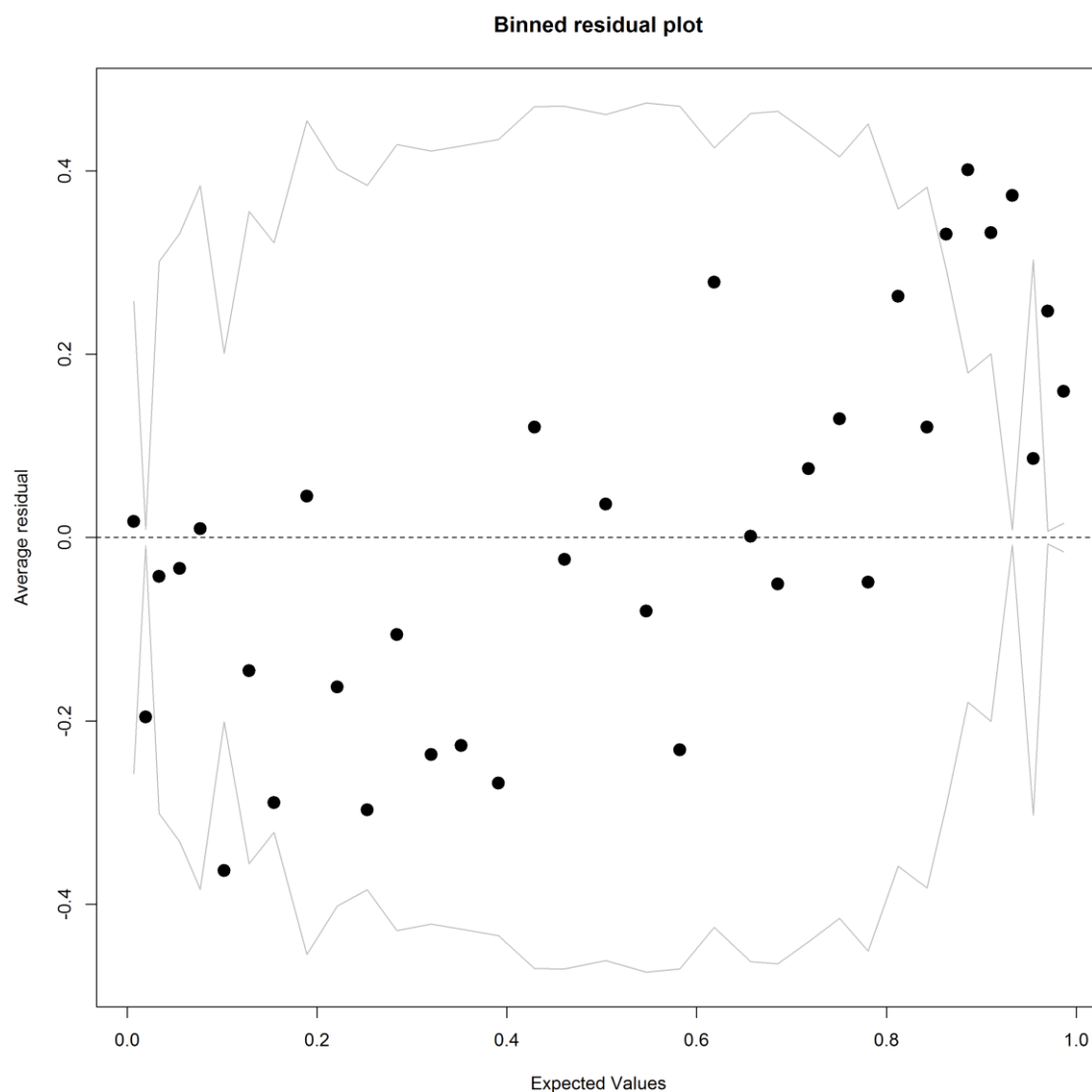


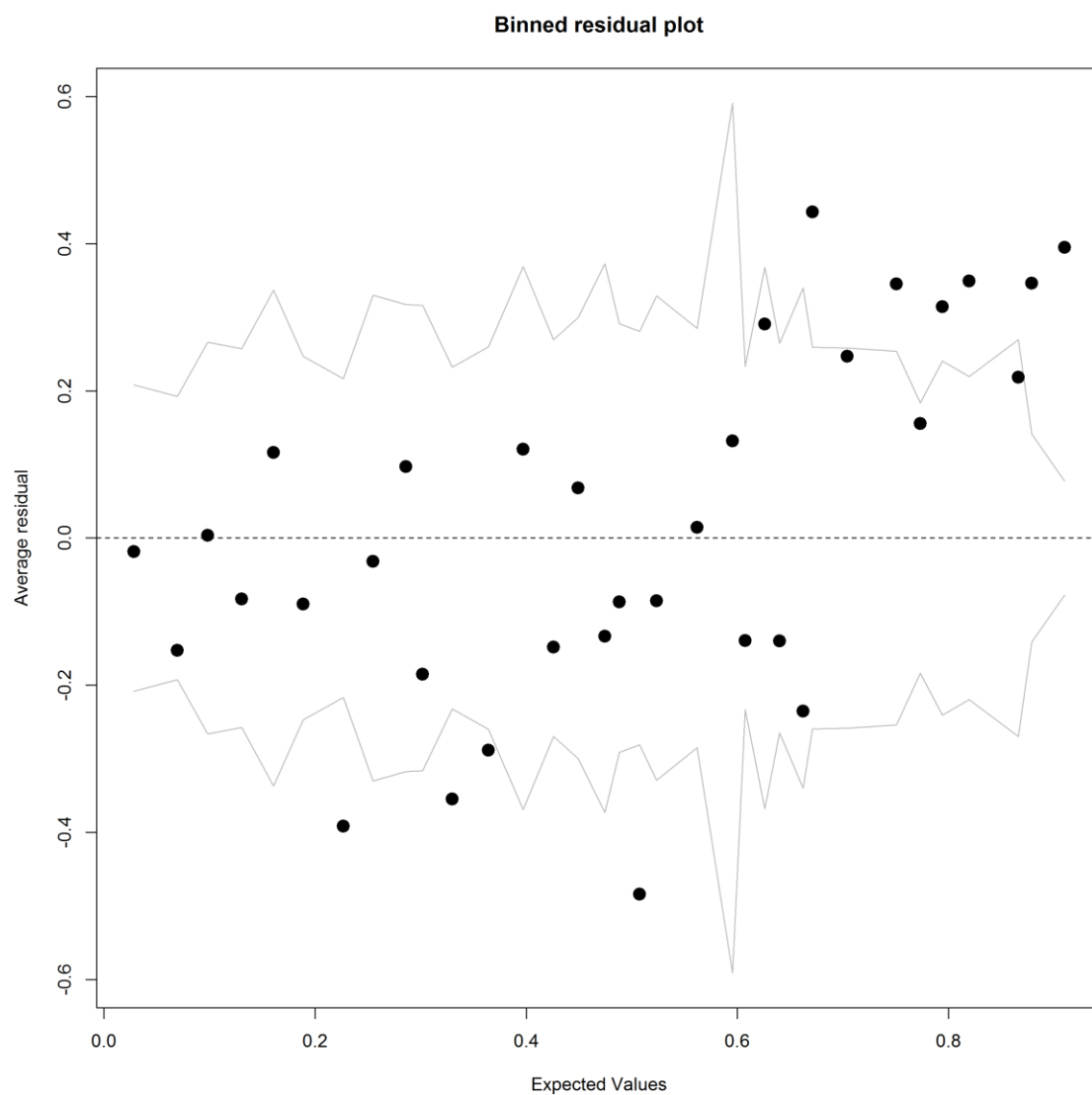
Figure 5-25. A residual plot of the ethanol model (from subset data no.1).

The raw residual plot of a logistic regression model isn't particularly informative. It cannot be estimated by the naked eye how many points fall above or below the curve, making the judgment of the presence of a trend or a bias not possible. A suggested way to go around this problem is to convert the data into bins (groups), where the average residual value among these can be seen. The binned residuals plot for an equal-size sample prediction for each model is shown in Figure 5-30 (a-e).



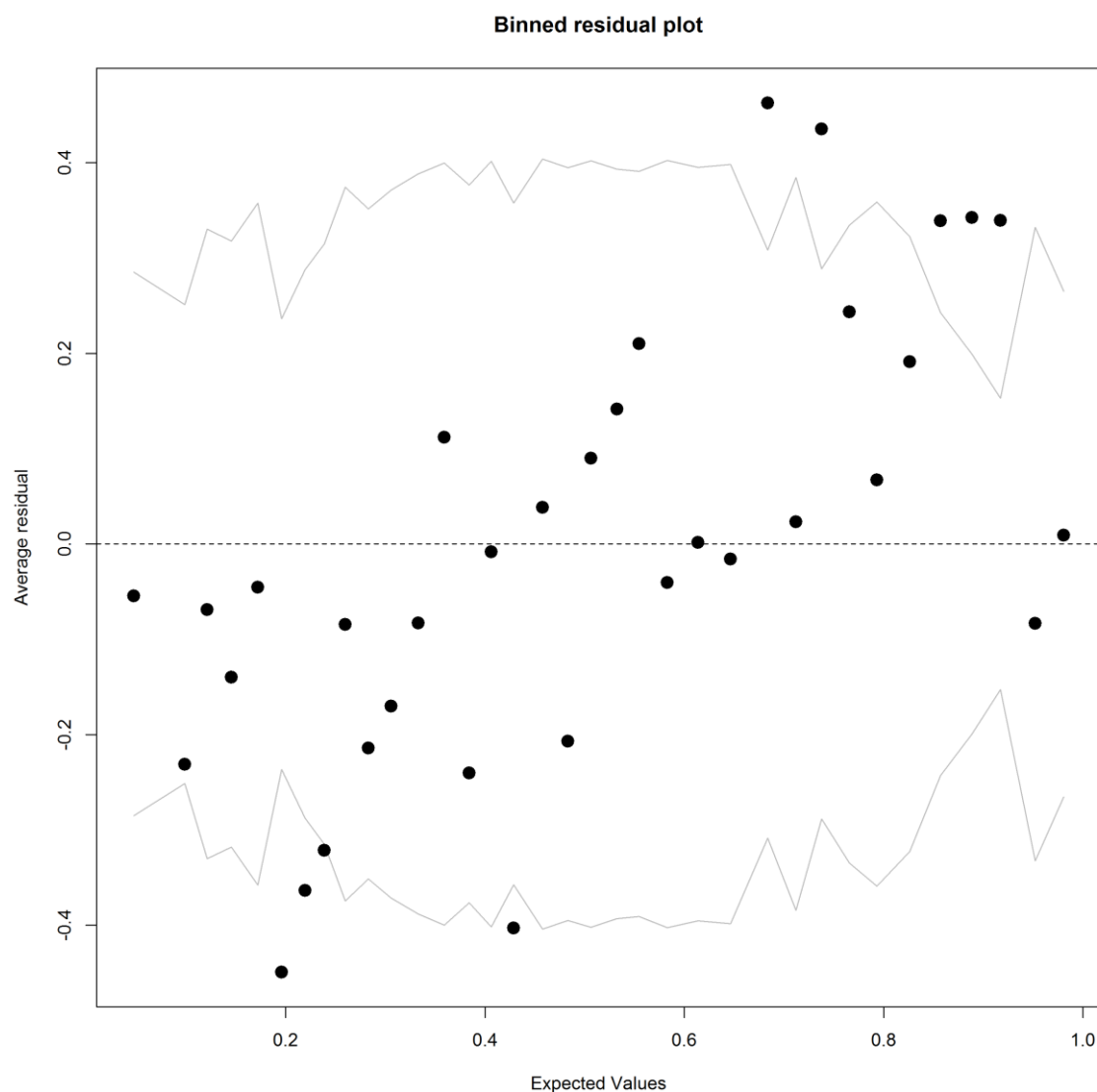
(a)

Figure 5-26. The binned residual plots from the 2 variable models in each solvent (a): ethanol. Note that each plot is based on an equal size sample to show sensible distribution around the zero line. The two grey lines in each figure represents the boundaries of 95 % error assuming the model is true.



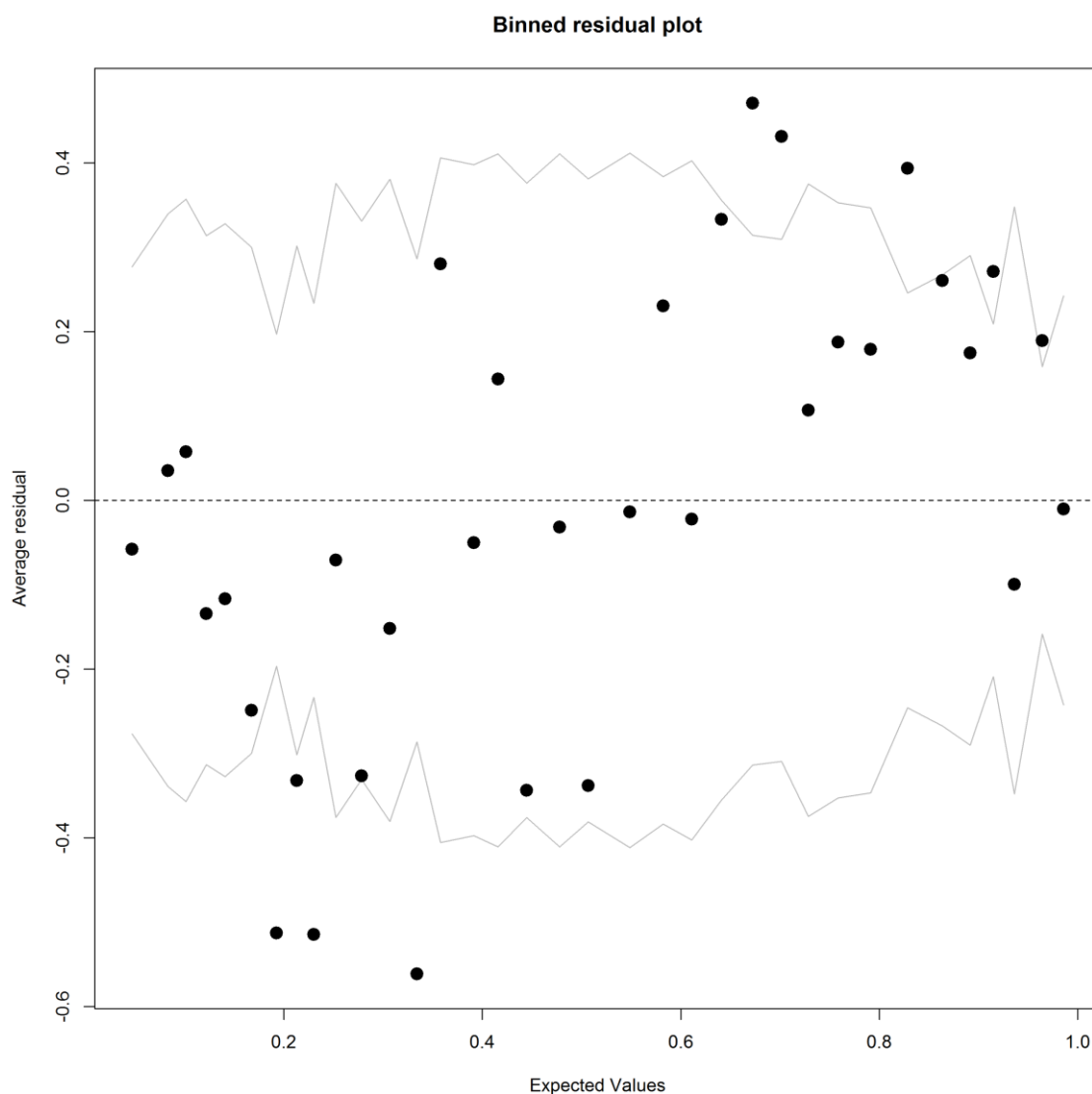
(b)

Figure 5-27. Continued. The binned residual plots from the 2 variable models in each solvent (b): methanol. Note that each plot is based on an equal size sample to show sensible distribution around the zero line. The two grey lines in each figure represents the boundaries of 95 % error assuming the model is true.



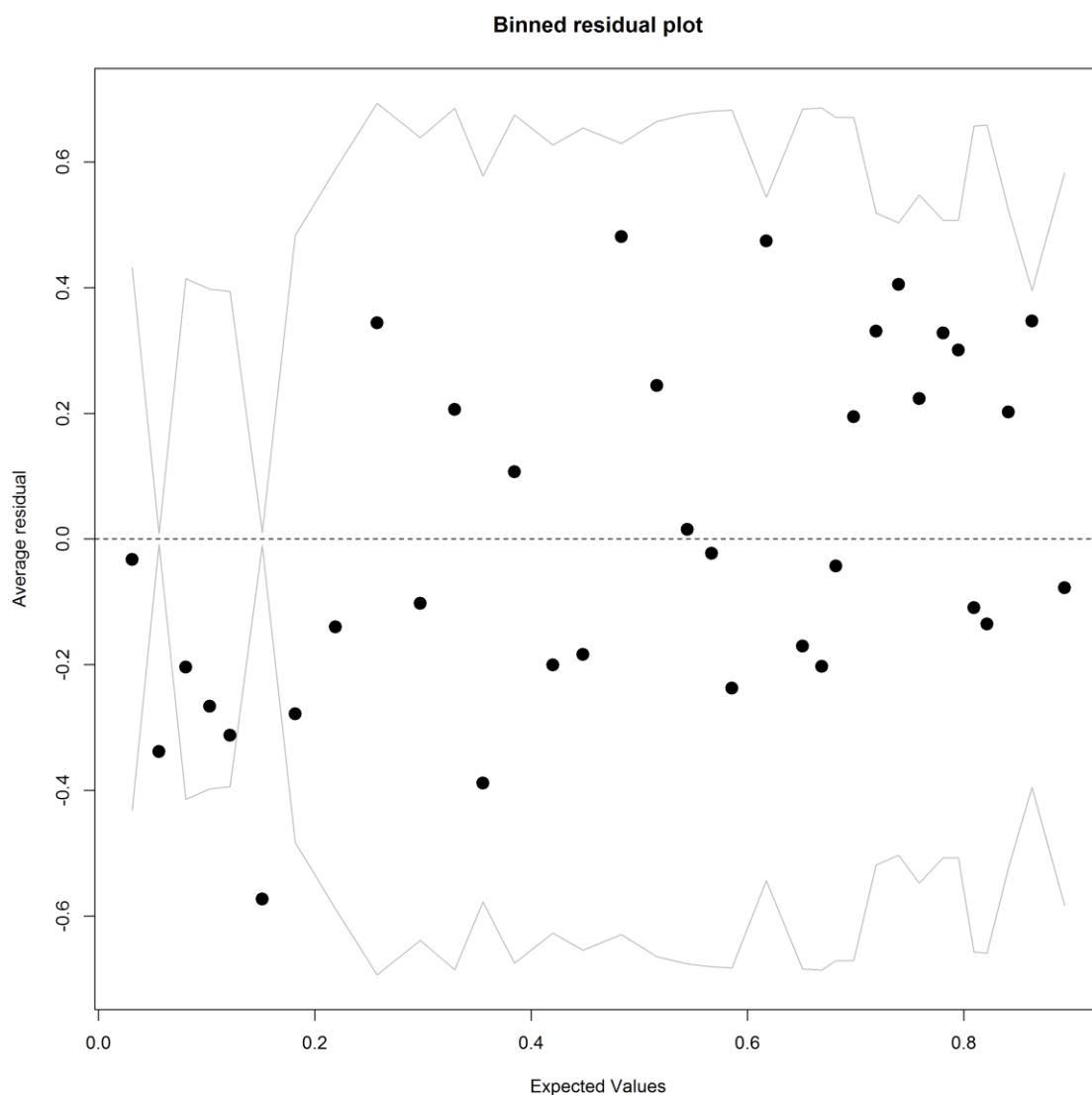
(c)

Figure 5-28. Continued. The binned residual plots from the 2 variable models in each solvent (c): dichloromethane. Note that each plot is based on an equal size sample to show sensible distribution around the zero line. The two grey lines in each figure represents the boundaries of 95 % error assuming the model is true.



(d)

Figure 5-29. Continued. The binned residual plots from the 2 variable models in each solvent (d): chloroform. Note that each plot is based on an equal size sample to show sensible distribution around the zero line. The two grey lines in each figure represents the boundaries of 95 % error assuming the model is true.



(e)

Figure 5-30. Continued. The binned residual plots from the 2 variable models in each solvent (e): water. Note that each plot is based on an equal size sample to show sensible distribution around the zero line. The two grey lines in each figure represents the boundaries of 95 % error assuming the model is true.

The residual plots for all models look normal, with no patterns observed. Although one sample is shown here from each solvent, 10 similar plots were exported for 10 different test sets for each solvent's model, where they showed similar performance.

5.6.7 Misclassified data and intercept adjustment

It was shown earlier that each solvent's model (the two-variable models in shown in section 5.5.1) misclassified between 19 and 26 % of the starting (complete) datasets. But was most misclassification attributed to solvate or non-solvate group or was it a 50-50 % split between them? Ideally, the misclassification should be a half-split between the solvate and the non-solvate groups. This is thought to be due to the fact that the cut-off point was chosen in a way to maximize the sensitivity and the specificity of the models, as has been shown in section 5.5.5 The percentage of solvates and non-solvates among the misclassified data in each solvent dataset was calculated. The results are shown in Figure 5-31.

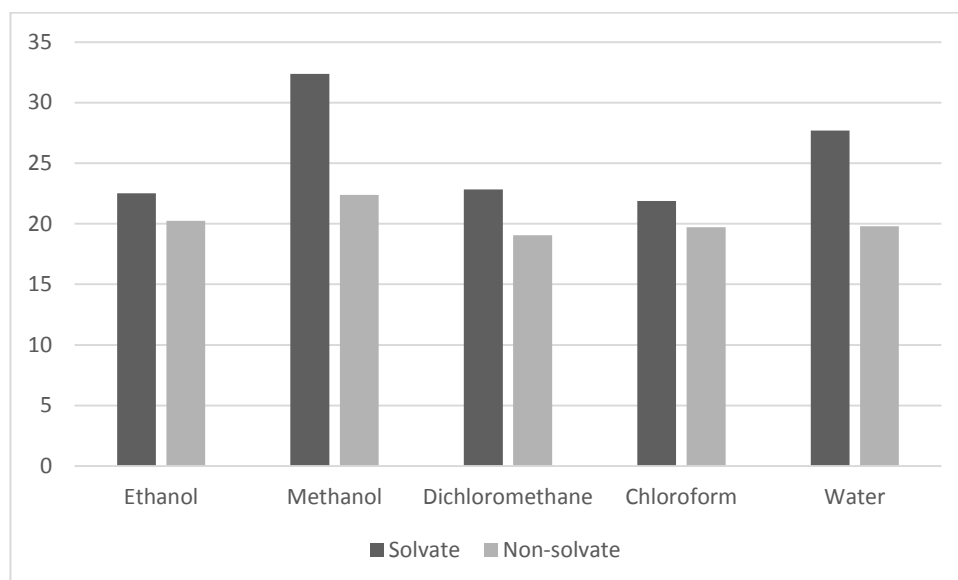


Figure 5-31. Percentage of misclassifications in the solvate and the non-solvate groups in each model when the complete dataset was predicted using the two-variable models.

All models have shown a biased behaviour towards the non-solvate group to different extents, with the alcohols and water showing the largest bias. This could be surprising in the first instance, but bearing in mind that the prior probability of the solvate formation was ignored when equal size samples were taken explains this. Although 10 subsets were used to fit each model, the fact that equal size sampling was used results in ignoring part of the data. In other

words, this error happened due to the sampling error. In a logistic regression model, this error could be fixed by adjusting the intercept of the model.²⁵ Therefore, what is required at this stage is eliminating the bias that is present in these models *via* intercept adjustment.

In order to know the perfect cut-off point for each model, a loop programmed in R was used to make a stepwise change to the intercept in both positive and negative direction until the bias in prediction between the solvate and the non-solvate groups is eliminated. This adjustment results in an equally distributed error between the solvate and the nonsolvate groups, as shown in Figure 5-32.

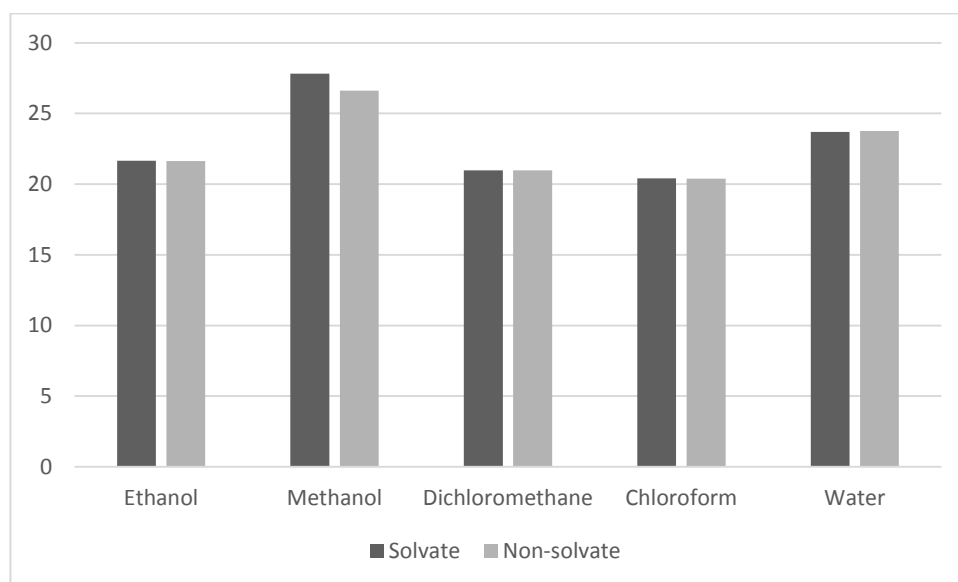


Figure 5-32. Percentage of misclassifications in the solvate and the non-solvate groups in each model when the complete dataset was predicted using the intercept-adjusted two-variable models.

Intercept-adjusted models are more fair towards both the solvate and non-solvate groups, nevertheless, the overall misprediction and MSE of the models has shown a small increase in some solvents. The change in intercept is shown in Table 5-13. The details on the change of the overall prediction and the MSE per solvent are shown in Figure 5-33.

Table 5-13. The intercept value in normal and intercept-adjusted models

	Original	Adjusted
Ethanol	15.939	15.868
Methanol	2.8	2.652
Dichloromethane	15.459	15.357
Chloroform	14.744	14.688
Water	4.672	4.558

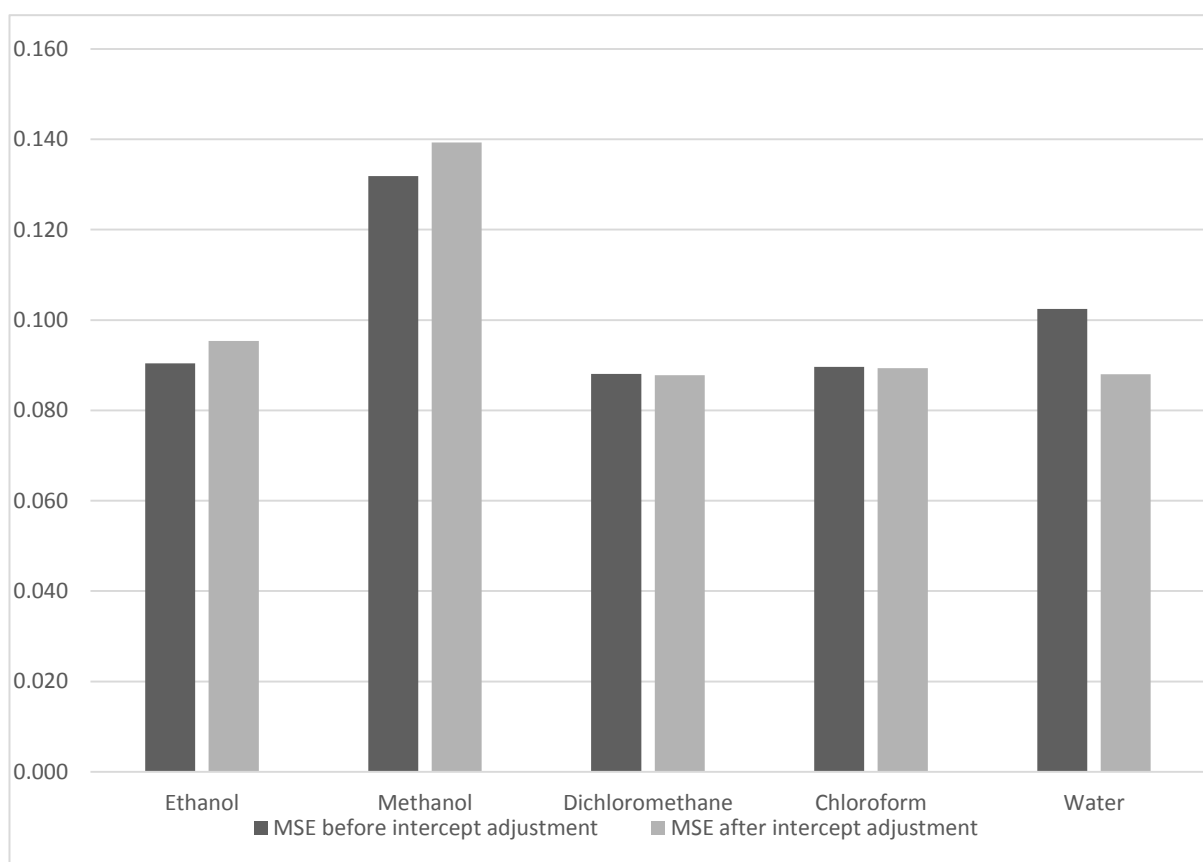


Figure 5-33. The effect of adjusting the intercept on the MSE value of each model.

5.7 Simple alternatives

To this point, it is well known that the two-variable models were able to correctly predict the behaviour of 74-80 % of the data in any of the five solvents. The descriptors in the models shown in section 5.6.1 can be instantly calculated by a computer for almost any molecule. Nevertheless, it would be impractical to calculate them manually. For example the process of

calculating the value of the SM3_H2 descriptor involves working out the third power of a matrix. This is a tedious task to be performed manually for almost any molecule. Additionally, the descriptor value is not easily estimated by looking at the molecular structure. In order to give a more intuitive value, simple alternative models were introduced.

5.7.1 The alternative descriptors

The rationale for solving this issue was based on finding simple models that resemble the original ones given in section 5.5.1. This could be achieved by considering for simple descriptors that are highly correlated with the ones in the original models. A correlation matrix was established for each solvent's dataset individually and new descriptors were selected.

In the ethanol model, the AVS_H2 descriptor requires a lot of time to be calculated. A correlation matrix was calculated and arranged according to the correlations between AVS_H2 and the rest of descriptors. Part of this correlation matrix is presented in Table 5-14.

Table 5-14. Part of the AVS_H2 correlation matrix

-	AVS_H2
AVS_H2	1
SpMax_H2	0.981
SpDiam_H2	0.977
SM6_H2	0.975
...	...
nCIC	0.870

The top 3 correlated descriptors in this dataset were SpMax_H2, SpDiam_H2 and SM6_H2. Although these had the highest correlation, the notation (H2) in their names indicate the need to calculate the reciprocal squared topological distance matrix, therefore they are not suitable to be manually calculated. The descriptor with 109th highest correlation was the nCIC at a correlation value of 0.87. This is a simple count descriptor that counts the number of rings in a molecule. The correlation sounds logical as AVS_H2 inherently contains information about the size and branching of a molecule; therefore a larger molecule is expected to have a higher number of rings. For this reason, the AVS_H2 was replaced with the nCIC and a model was fitted. The second descriptor in the ethanol model was already a simple count descriptor, so it was not changed. The average MSE of the model that uses nCIC and nHDon over 10 samples was 0.157, which is close to 0.148, the average MSE of the original model.

In the methanol model, the descriptors used were the TRS and nHDon. Both of these are simple descriptors that are easy to calculate manually, for this reason the methanol model was not adjusted or represented by a simpler model.

In the dichloromethane model, the SM3_H2 descriptor hard to be estimated using the molecular graph. Fortunately, a simple path count descriptor (*MPC01*) showed to be very similar to *SM3_H2*, with a high correlation ($r = 0.983$). *MPC01* is the count of paths of length 1 in the H-depleted molecular graph. It could also be seen as the number of bonds between non-hydrogen atoms in the molecular graph.^{26, 27} Just like *SM3_H2*, *MPC01* shows large values when the molecule size increases so it was subject to logarithmic transformation.

The second descriptor in the dichloromethane model was the Hy. This descriptor is not very complex but may require a considerable amount of time to be calculated. For this reason a simpler alternative was looked up. nHDon turned out to have a correlation of more than 0.95 to Hy, so it was replaced by nHDon. The resulting model still gives similar results. This

alternative simple model had an average MSE of 0.150 over 10 samples compared to 0.146 for the original model.

In the chloroform model, the first descriptor (SM3_H2) was shared with dichloromethane. Interestingly, chloroform also shares the first alternative descriptor with dichloromethane, where the MPC01 descriptor showed a correlation of 0.984 to SM3_H2. The second descriptor in the chloroform model was the number of hydrogen atoms attached to a heteroatom (H-050). This is a simple count descriptor, therefore it was left unchanged. Although the chloroform model shows a different second descriptor from the dichloromethane model, the correlation between the second descriptors in both is very high ($r > 0.95$ correlation), indicating the similarity in their behaviour. The average MSE of the simpler model, again over 10 samples, is 0.152, compared to 0.148 of the original model.

In the water model, the first descriptor was the π ID. This descriptor is calculated through the formula presented in Equation (5-8). Since it is not easy to calculate, it was replaced with the nCIC descriptor, which is the number of rings in a molecule. nCIC has a correlation value ($r = 0.854$) of with π ID. The second descriptor in the water model was the nH, which is already a simple count descriptor. Therefore it was left as it is. A model utilizing the nCIC and nH descriptors has an average MSE of 0.161 compared to an average MSE of 0.159 of the original hydrate model. The result of these replacement was 5 new models. These five models had their intercept adjusted in the same manner shown in section 5.6.7. The resulting models are illustrated in Equations (5-11) to (5-15):

$$p_{ethanol} = 1 - \frac{1}{1 + e^{-(3.952 - 0.766nCIC - 0.889nHDon)}} \quad (5-11)$$

$$p_{methanol} = 1 - \frac{1}{1 + e^{-(2.652 - 0.085TRS - 0.612nHDon)}} \quad (5-12)$$

$$p_{dichloromethane} = 1 - \frac{1}{1 + e^{-(12.737 - 3.649MPC01 - 0.339nHDon)}} \quad (5-13)$$

$$p_{chloroform} = 1 - \frac{1}{1 + e^{-(12.336 - 3.416MPC01 - 0.358H.050)}} \quad (5-14)$$

$$p_{water} = 1 - \frac{1}{1 + e^{-(2.45 - 0.606nCIC - 0.088nH)}} \quad (5-15)$$

5.7.2 Performance of the simple models

The descriptors chosen to simplify the models showed a high correlation to the ones in the original models as has been shown in section 5.6.1. This means the simple alternatives are ought to have a performance that is comparable to the original ones. The average performance of the alternative models (over 10 samples) is shown in Table 5-15. A comparison of the MSE between the original and the alternative models is shown in Figure 5-34.

Table 5-15. The average performance of the intercept-adjusted alternative models (over 10 samples) in each solvent's dataset.

Model	Descriptors	Intercept	Descriptor 1 coefficient	Descriptor 2 coefficient	No. of datapoints	MSE	AIC	AUC
Ethanol	<i>nCIC</i> + <i>nHDon</i>	3.952	-0.766	-0.889	1377	0.157	1320	0.853
Methanol	<i>TRS</i> + <i>nHDon</i>	2.808	-0.084	-0.612	3035	0.180	3268	0.810
Dichloromethane	<i>MPC01</i> + <i>nHDon</i>	13.236	-3.649	-0.339	2592	0.150	2428	0.865
Chloroform	<i>MPC01</i> + <i>H-050</i>	12.416	-3.416	-0.358	2384	0.152	2254	0.861
Water	<i>nCIC</i> + <i>nH</i>	2.45	-0.606	-0.088	607	0.165	597	0.835

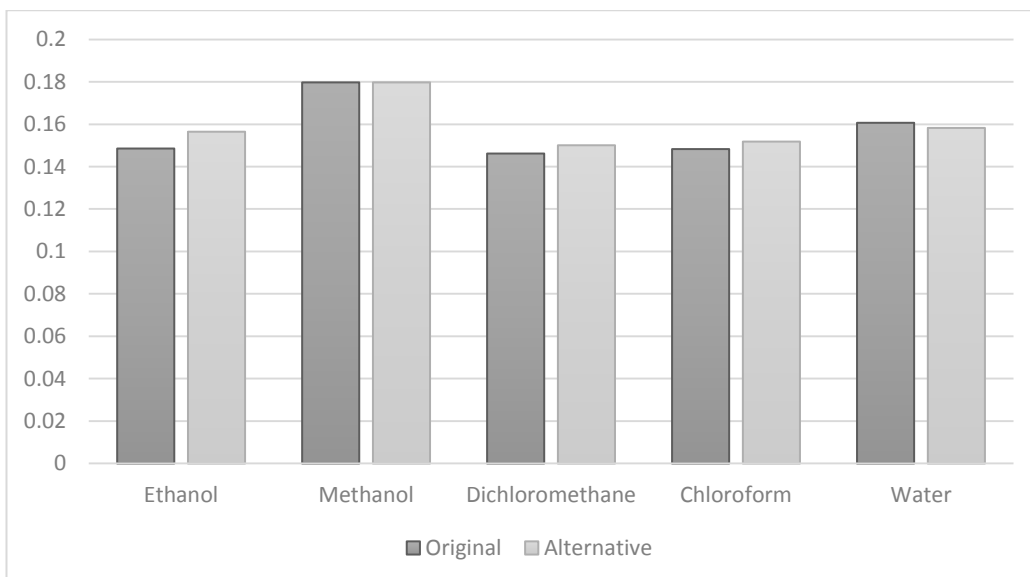


Figure 5-34. The increase in the MSE value when alternative models are used. Note that the comparison in shown here is between intercept-adjusted two-variable models and intercept-adjusted alternative models.

The increase in the error values was minimal. This demonstrates that predictions with the same success rate can be made without the use of a computer.

5.8 Practical usage of the models

The application of the models can be viewed in two ways, these are a visual representation and a purely mathematical one, both of which are going to be explained in this section.

5.8.1 Mathematical representation

After the alternative models were introduced, it is logical to show an example of how these could be used to make a prediction. Bearing in mind that most users would not have the descriptor calculation software, it was necessary to demonstrate their practical usage. In this section, two molecules are going to be predicted for solvate formation using the alternative models. The molecules chosen to exemplify the usage of the models were large, branched and complex. The reason for these choices is to demonstrate the easiness of using the alternative models, regardless of the complexity of the structure. The first example is going to utilize the

water alternative model. The example of choice was the azithromycin, an anti-bacterial from the macrolides family (CSD reference code: NAVTAF28), which is shown in Figure 5-35.

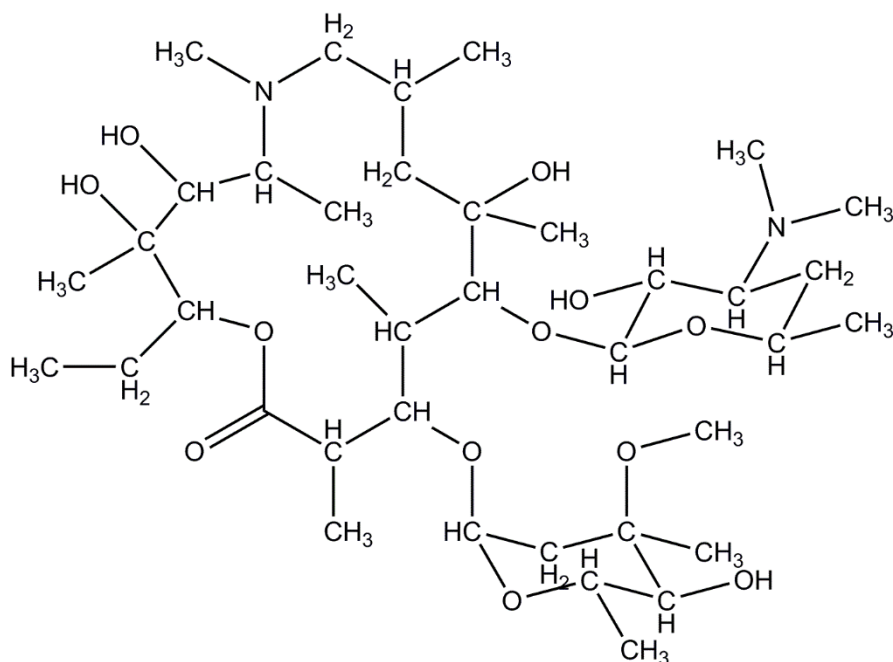


Figure 5-35. The structure of azithromycin. Chemical formula $C_{37}H_{67}NO_{13}$.

In order to calculate the probability of hydrate formation, two descriptors are required, these are the nCIC and nH. The nCIC (number of rings) in the structure can be easily worked out as 3, while the nH (number of hydrogen atoms) would take a little longer. The number of H atoms sum up to 67. Putting these in the alternative water model gives the Equation (5-16):

$$p_{water} = 1 - \frac{1}{1 + e^{-(2.45 - 0.606 \cdot 3 - 0.088 \cdot 72)}} = 0.003 \quad (5-16)$$

The prediction is well below the cut-off point for the water model, which indicates this molecule is a strong candidate for forming a hydrate. In reality, a dihydrate is recorded in the CSD for this entry (GEGJAD), which makes the prediction correct in this case.

Another example could be demonstrated *via* the methanol alternative model. This time, a pharmaceutically active statin, known as bryostatin was chosen to predict solvate formation. Similar to the previous example, the values of two descriptors have to be worked out in order to make the prediction. This time the two descriptors were TRS and nHDon. Considering the structure of the molecule, the TRS descriptor can be calculated by counting the number of atoms in every ring. This structure show 4 rings, 3 of which are 6-membered rings and one that is a 20-membered ring. These rings overlap, but the number of atoms consisting each ring are counted individually, leaving the final value of the TRS descriptor to be 38. The nHDon descriptor value for this structure is 4. An illustration of the two descriptors on the bryostatin structure is shown in Figure 5-36.

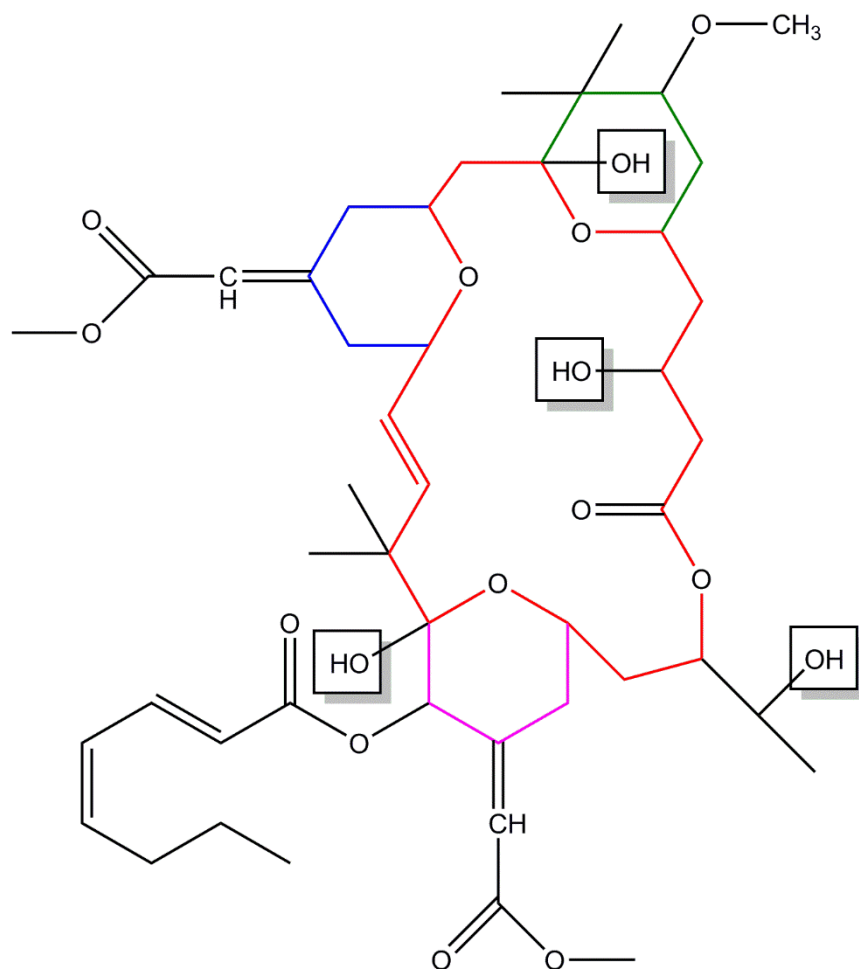


Figure 5-36. The chemical structure of bryostatin with the groups contributing to the TRS and nHDon descriptors highlighted.

Feeding these values into the model gives the result shown in Equation (5-17):

$$p_{\text{methanol}} = \frac{1}{1 + e^{-(2.652 - 0.0847 \cdot 38 - 0.612 \cdot 4)}} = 0.047 \quad (5-17)$$

The low probability value gives the indication of a high likelihood of the methanol solvate formation. In fact, this molecule does form a methanol solvate. The form was reported in a publication in 1982 (CSD reference code: BOKKIV²⁹).

5.8.2 Visual representation

From a visual perspective, the two-variable models are representing a two-dimensional problem, as has been shown in Figure 5-21. In this problem, the two descriptors will represent the two axes of a plane. For each given molecule, the values of these two descriptors are calculated. These two values are used as the coordinates to draw a point on the plane. The machine then draws the decision boundary (which is defined by the model) on the same plane. The position of the drawn point in relation to the decision boundary decides the prediction outcome. The distance from this decision boundary defines the likelihood of hydrate or solvate formation. An example is given here using the amoxicillin molecule, which was described back in 1971.³⁰ An illustration of the amoxicillin molecule and its classification by the model are shown for the amoxicillin molecule in Figure 5-37 and Figure 5-38, respectively.

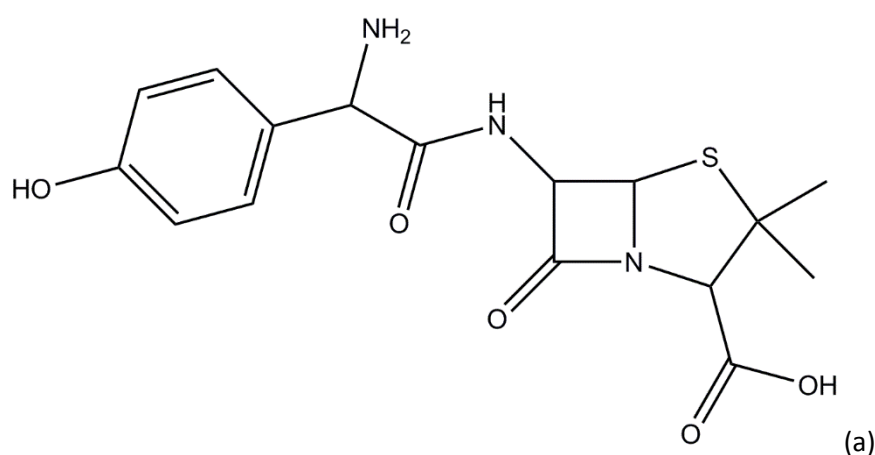


Figure 5-37. The amoxicillin molecule.

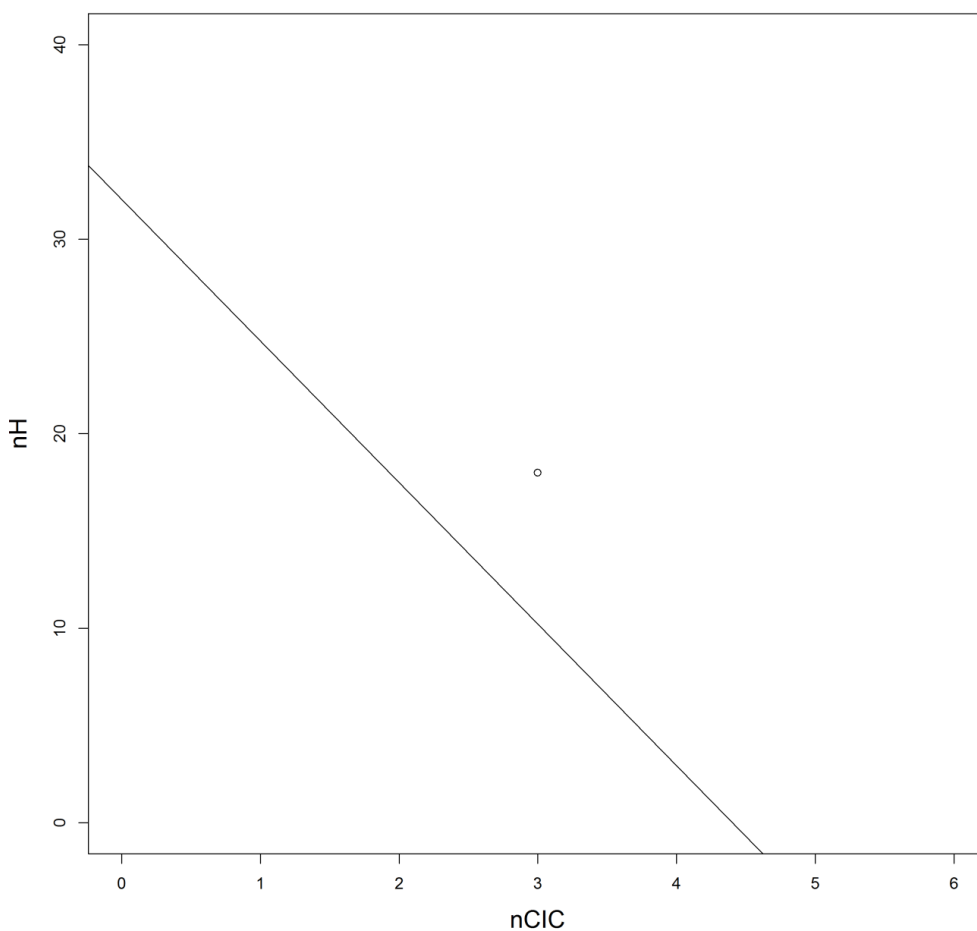


Figure 5-38. An illustration of the graphical prediction method used by the water model. The point in the plot represents the amoxicillin molecule.

By looking at the water model shown in Equation (5-15), it can be anticipated that the region for hydrate formation lies to the right of the decision boundary. The amoxicillin did fall into this region, with the value of the two descriptors being 3 and 18, respectively. This indicates the drug's ability to form a hydrate according to the model. This is a correct prediction as the amoxicillin drug was found to have a trihydrate form that was recognized in the late 1970s.³¹ The trihydrate entry is recorded in the CSD under the reference code AMOXCT10.

5.9 References

1. Dahiru T. P-value, a true test of statistical significance? A cautionary note. *Annals of Ibadan postgraduate medicine*. 2008;6(1):21-6.
2. Cordella CBY. PCA: The basic building block of chemometrics: INTECH Open Access Publisher; 2012.
3. Hsu C-W, Chang C-C, Lin C-J. A practical guide to support vector classification. 2003
4. Berthold MR, Cebron N, Dill F, Gabriel TR, Kötter T, Meinl T, et al., editors. KNIME: The Konstanz Information Miner. *Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007)*; 2007: Springer.
5. Kotsiantis SB, Zaharakis ID, Pintelas PE. Machine learning: a review of classification and combining techniques. *Artificial Intelligence Review*. 2006;26(3):159-90.
6. Sedgwick P. What is sampling error? *BMJ : British Medical Journal*. 2012;344.
7. Gaspar P, Carbonell J, Oliveira JL. On the parameter optimization of Support Vector Machines for binary classification. *Journal of Integrative Bioinformatics*. 2012;9(3):201.
8. Huang C-L, Wang C-J. A GA-based feature selection and parameters optimization for support vector machines. *Expert Systems with applications*. 2006;31(2):231-40.
9. Canty A, Ripley B. boot: Bootstrap R (S-Plus) Functions. R package version 1.3–11. Vienna: R Foundation for Statistical Computing. 2014.

10. Davison AC, Hinkley DV. Bootstrap methods and their application: Cambridge university press; 1997.
11. Analytics R, Weston S. Foreach: provides foreach looping construct for R. R package version. 2015.
12. Everitt B. Cambridge dictionary of statistics: Cambridge University Press; 1998.
13. Mihalić Z, Trinajstić N. A graph-theoretical approach to structure-property relationships. Journal of Chemical Education. 1992;69(9):701.
14. Estrada E. Spectral Moments of the Edge Adjacency Matrix in Molecular Graphs. 1. Definition and Applications to the Prediction of Physical Properties of Alkanes. Journal of Chemical Information and Computer Sciences. 1996;36(4):844-9.
15. Todeschini R, Vighi M, Finizio A, Gramatica P. 3D-modelling and prediction by WHIM descriptors. Part 8. Toxicity and physico-chemical properties of environmental priority chemicals by 2D-TI and 3D-WHIM descriptors. SAR and QSAR in Environmental Research. 1997;7(1-4):173-93.
16. Viswanadhan VN, Ghose AK, Revankar GR, Robins RK. Atomic physicochemical parameters for three dimensional structure directed quantitative structure-activity relationships. 4. Additional parameters for hydrophobic and dispersive interactions and their application for an automated superposition of certain naturally occurring nucleoside antibiotics. Journal of Chemical Information and Modeling. 1989;29(3):163-72.

17. Ghose AK, Viswanadhan VN, Wendoloski JJ. Prediction of Hydrophobic (Lipophilic) Properties of Small Organic Molecules Using Fragmental Methods: An Analysis of ALOGP and CLOGP Methods. *The Journal of Physical Chemistry A*. 1998;102(21):3762-72.
18. Randić M, Jurs PC. On a Fragment Approach to Structure-activity Correlations. *Quantitative Structure-Activity Relationships*. 1989;8(1):39-48.
19. Gasteiger J, Sadowski J, Schuur J, Selzer P, Steinhauer L, Steinhauer V. Chemical Information in 3D Space. *Journal of Chemical Information and Modeling*. 1996;36(5):1030-7.
20. Jetti RKR, Griesser UJ, Krivovichev S, Kahlenberg V, Blaser D, Boese R. Supramolecular synthesis of caffeine solvates and cocrystals. *Acta Crystallographica Section A*. 2005;61(1):286.
21. Harmon KM, Webb AC. Hydrogen bonding: 70. Thermodynamic and infrared study of stability and stoichiometry of tetramethonium and trimethonium bromide and chloride hydrates. *Journal of Molecular Structure*. 1999;508(1-3):119-28.
22. Infantes L, Fábián L, Motherwell WDS. Organic crystal hydrates: what are the important factors for formation. *CrystEngComm*. 2007;9(1):65-71.
23. Carter JV, Pan J, Rai SN, Galandiuk S. ROC-ing along: Evaluation and interpretation of receiver operating characteristic curves. *Surgery*. 2016;159(6):1638-45.
24. Gelman A, Hill J. Data analysis using regression and multilevel/hierarchical models: Cambridge University Press; 2006.
25. Owen AB. Infinitely imbalanced logistic regression. *Journal of Machine Learning Research*. 2007;8(Apr):761-73.

26. M. Randić. Characterization of Atoms, Molecules and Classes of Molecules Based on Paths Enumerations. *MATCH Communications in Mathematical and in Computer Chemistry*. 1979;7:5-64.
27. Randić M, Brissey GM, Spencer RB, Wilkins CL. Search for all self-avoiding paths for molecular graphs. *Computers & Chemistry*. 1979;3(1):5-13.
28. Stephenson GA, Stowell JG, Toma PH, Pfeiffer RR, Byrn SR. Solid-state investigations of erythromycin a dihydrate: Structure, NMR spectroscopy, and hygroscopicity. *Journal of Pharmaceutical Sciences*. 1997;86(11):1239-44.
29. Pettit GR, Herald CL, Doubek DL, Herald DL, Arnold E, Clardy J. Isolation and structure of bryostatin 1. *Journal of the American Chemical Society*. 1982;104(24):6846-8.
30. Long AAW, Nayler JHC, Smith H, Taylor T, Ward N. Derivatives of 6-aminopenicillanic acid. Part XI. α -Amino-p-hydroxy-benzylpenicillin. *Journal of the Chemical Society C: Organic*. 1971(0):1920-2.
31. Boles MO, Girven RJ, Gane PAC. The structure of amoxycillin trihydrate and a comparison with the structures of ampicillin. *Acta Crystallographica Section B*. 1978;34(2):461-6.

Chapter 6: Discussion of the models

The previous chapter has covered the statistical aspects of the models. In this chapter, we discuss examples from the dataset, representing the importance of the descriptors that were included in the models. Additionally, some descriptors that could be useful for determining solvate formation but were not part of the models will be discussed and represented by examples from the dataset. Such criticism can help the identification of the strong and weak points of the models. An important term to point out before the chapter starts is the short contact. In this work, this term is used to describe the distance between two atoms when it is at least 0.1 Å shorter than the sum of the van der Waals radii of the atoms in contact.

6.1 Effects the models take into account

At this point, it has been shown that the best descriptors to predict solvate formation were related to the size and branching of a molecule in addition to its hydrogen bonding ability. The relative importance of these two variables will be illustrated *via* examples from the ethanol datasets.

6.1.1 Size and branching

This molecular property is the most important in the prediction of the solvate formation according to the models. The significance of this property can be illustrated by discussing entries that have a small size, low branching and a high ability to form hydrogen bonds and by discussing the other extreme case of large molecules with a little ability to form hydrogen bonds. A couple of examples representing the former case are 1-((E)-2-pyridinylmethylidene)semicarbazone, CSD reference code: KUHGEA¹ and N-(pyridin-2-yl)hydrazinecarbothioamide, CSD reference code: XAPTOY.² These are shown in Figure 6-1.

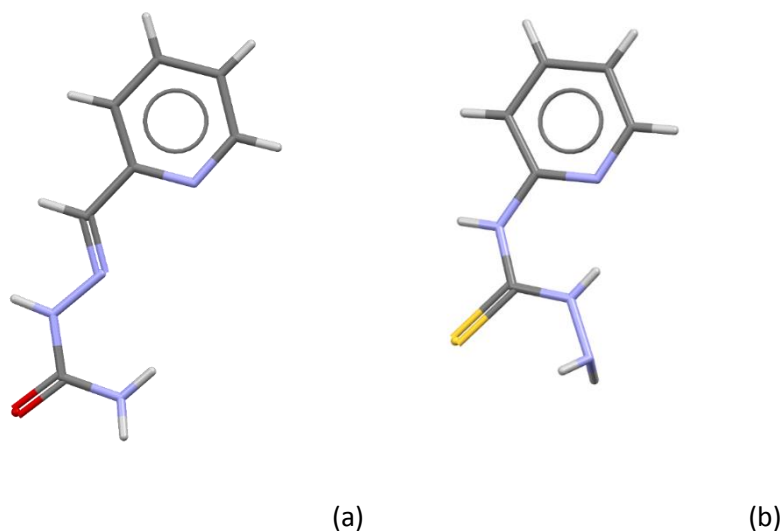


Figure 6-1. Molecular structures of (a) KUHGEA and (b) XAPTOY.

KUHGEA and XAPTOY possess AVS_H2 values of 2.972 and 2.979 and a number of hydrogen bond donors of 3 and 4, respectively. Both of these entries were recrystallized from ethanol. The solvent was not incorporated into the crystal structure in either of the cases, despite the availability of multiple hydrogen bond donors as well as acceptors. The inability of these entries to form an ethanol solvate was correctly predicted by the ethanol model at an x value of 0.874 for KUHGEA and 0.741 for XAPTOY. This inability to form a solvate was reasoned by not having a size and branching that is large enough to surpass the decision boundary into the solvate region. An illustrative example of the case of the molecules that show a large, branched structure can be the compound (16S)-(-)-16-ethylrhazilinam ethanol solvate, CSD reference code: ECEKON.³ This molecule shows a branched structure (AVS_H2 value is 4.099) and has one hydrogen bond donor (nHDon value is 1). The entry is an ethanol solvate, nevertheless, no hydrogen bond seems to exist between the molecule and the solvent even in the range of the Van der Waals radii + 0.1 Å. This shows that the hydrogen bond is not highly involved in the inclusion of the solvent inside the crystal structure. An illustration of the ECEKON entry is shown in Figure 6-2. Its solvate formation was correctly predicted by the ethanol model at $x = 0.345$.

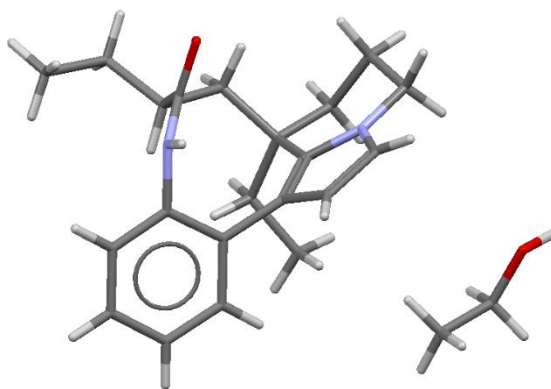


Figure 6-2. Molecular structures of ECEKON.

Multiple examples illustrating solvate formation despite the low count of hydrogen bond donors can be seen in all datasets. The CSD reference codes for additional examples from the ethanol and methanol datasets are given in Table 6-1.

Table 6-1. Additional examples on the importance of size and branching in solvate formation

From ethanol data				
CSD reference code	AVS_H2	nHDon	Chemical formula	Prediction value
MOKVOY ⁴	4.233	0		0.428

Table 6-1. Continued

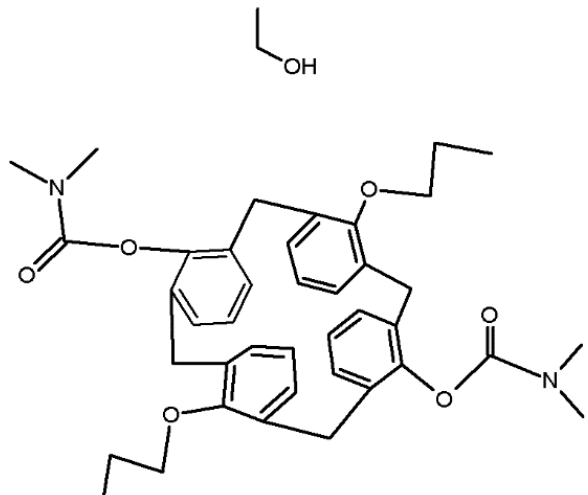
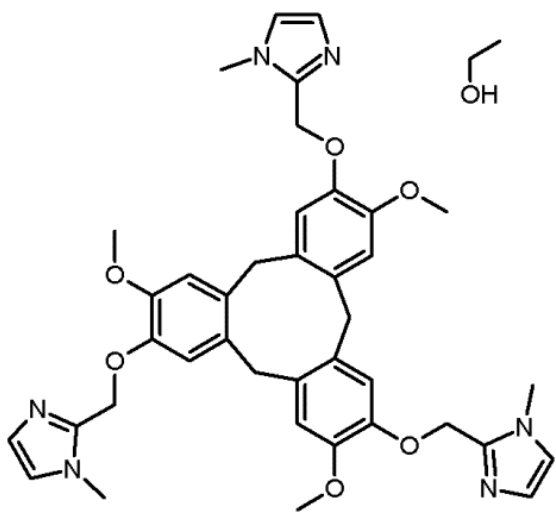
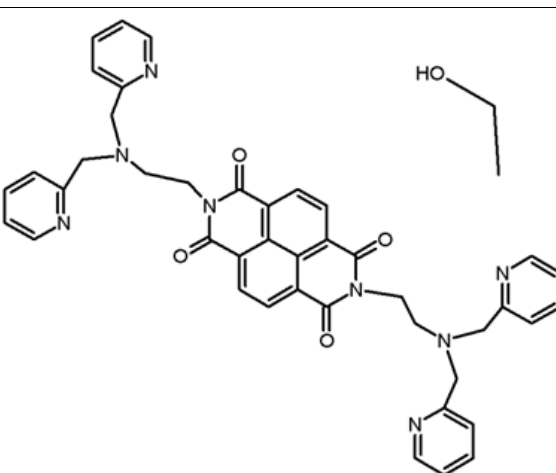
QEKMEY ⁵	4.201	0		0.458
QUSWUX ⁶	4.201	0		0.458
RIBBOU ⁷	4.213	0		0.447

Table 6-1. Continued

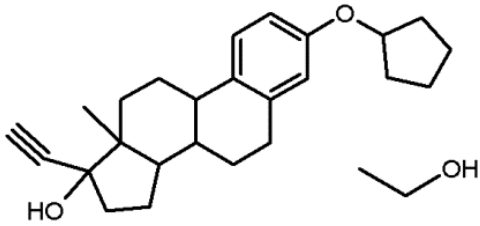
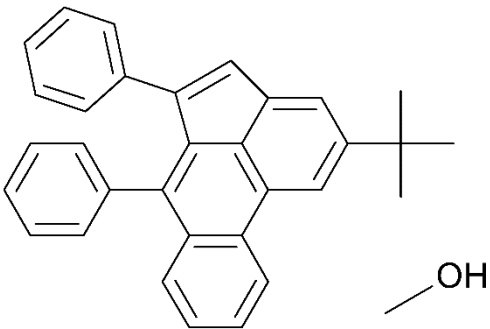
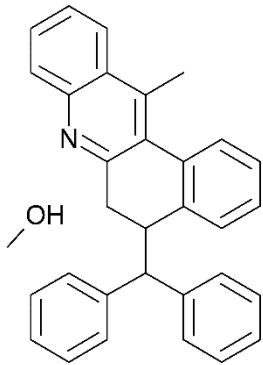
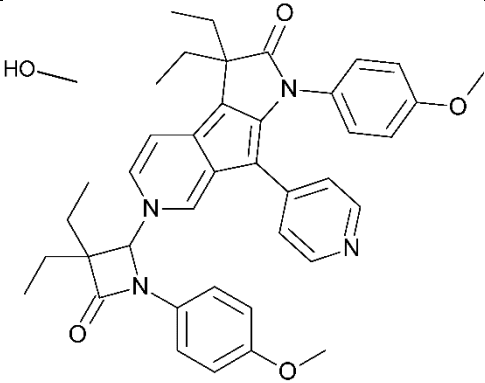
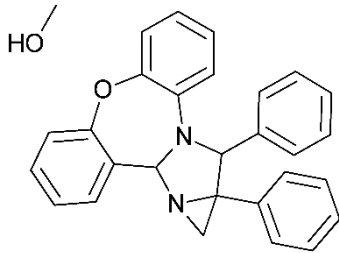
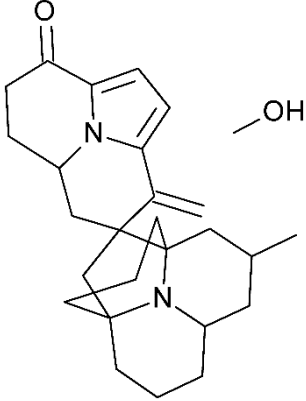
TOYJUM ⁸	4.092	1		0.352
From methanol data				
EREWED ⁹	35	0		0.420
EMEBUT ¹⁰	36	0		0.420
KAQVIJ ¹¹	38	0		0.359

Table 6-1. Continued

DADWEM ¹²	39	0		0.340
ZEBXUZ ¹³	40	0		0.321

6.1.2 Hydrogen bonding

The importance of hydrogen bonding in different structural formations, including solvate structures cannot be denied. Several studies have shown a positive correlation between the hydrogen bonding ability and the formation of solvates, especially in alcohol containing solvents. Additionally, all the models found agreed that the hydrogen bonding ability of the structure is an important factor in the determination of solvate formation. A number of examples from the ethanol dataset are used to demonstrate the importance of this type of bonding. The molecules that are used to make the demonstration were chosen to have similar AVS_H2 values, but differ in the number of hydrogen bond donors as illustrated in Figure 6-3 (a) and (b).

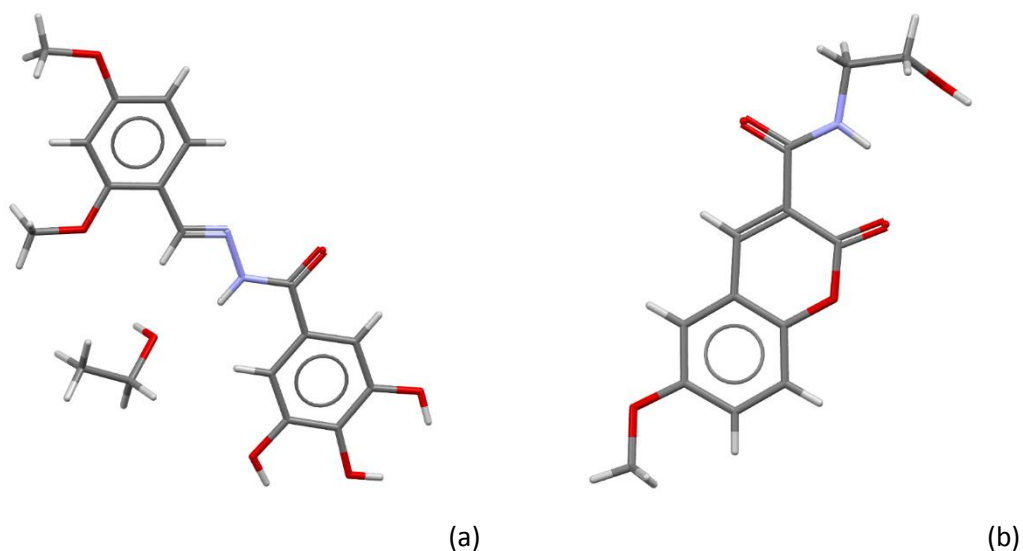


Figure 6-3. Molecular structures of (a) SOYQON and (b) UHUWEA.

The first example, shown in Figure 6-3(a) is N'-(2,4-dimethoxybenzylidene)-3,4,5-trihydroxybenzohydrazide ethanol solvate, CSD reference code: SOYQON.¹⁴ The other entry shown in Figure 6-3(b) is N-(2-hydroxyethyl)-6-methoxy-2-oxo-2H-chromene-3-carboxamide, CSD reference code: UHUWEA.¹⁵ Both compounds were recrystallized from ethanol, but the latter has failed to form a solvate. These two entries have close AVS_H2 values of 3.534 and 3.513, respectively. The larger difference between them lies in the hydrogen bonding ability, where the SOYQON entry had 4 hydrogen bond donors, compared to 2 donors in the UHUWEA entry. According to the model, this difference in the number hydrogen bond donors is the reason for the solvate formation in the first molecule. Both of these entries were correctly predicted for solvate formation by the ethanol model, where SOYQON had an x value of 0.256 and the UHUWEA entry had an x value of 0.676.

Another example can be shown *via* ethyl 5-((2-hydroxybenzoyl)carbohydrazonoyl)-3,4-dimethyl-1H-pyrrole-2-carboxylate ethanol solvate, CSD reference code: EHUKAU,¹⁶ and (E)-N'-(4-(2-chlorobenzyloxy)benzylidene)isonicotinohydrazide, CSD reference code: QERJED,¹⁷ both of which are illustrated in Figure 6-4.

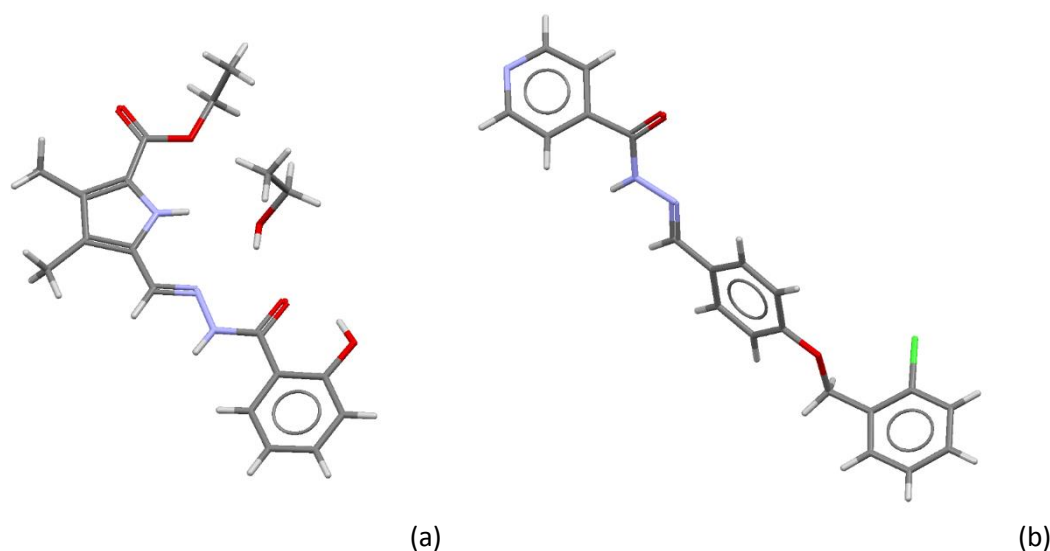


Figure 6-4. The importance of hydrogen bonding in solvate formation shown *via* the EHUKAU (a) and the QERJED (b) entries.

Both of these entries have an AVS_H2 value of 3.533. The count of the hydrogen bond donors is different though. While the EHUKAU entry shows three hydrogen bond donors, the QERJED has only one. This causes the model to predict them differently. Both of these entries were correctly predicted by the ethanol model, where the EHUKAU entry was predicted to form a solvate at $x = 0.450$ and the QERJED entry was predicted not to form a solvate at $x = 0.821$. A large number of examples were available to demonstrate the case. Additional paired examples illustrating this case are shown in Table 6-2. Note that these examples are from the ethanol and the methanol datasets only. This is because the models of these two solvents are the only ones that were significantly improved by the addition of the hydrogen bonding factor (see section 5.5.2 for the addition of the second variable).

Table 6-2. Additional paired examples indicating the importance of hydrogen bonding ability in solvate formation. Each pair of examples has similar background shading

From ethanol data

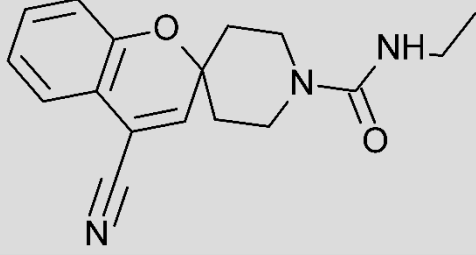
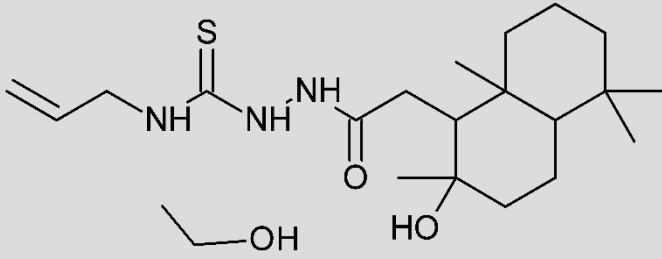
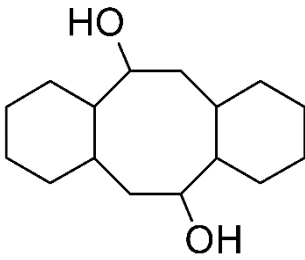
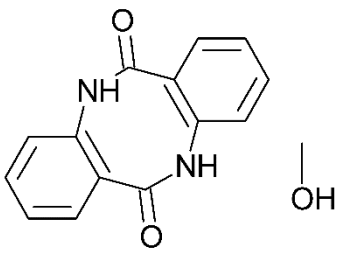
Chemical formula and reference code	AVS_H2 value	nHDon value	Prediction value
<p>NEQDEU¹⁸</p> 	3.743	1	0.673
<p>KAYTOU¹⁹</p> 	3.743	4	0.134
<p>AZUYAV²⁰</p> 	3.758	0	0.821
<p>ABEBOZ²¹</p> 	3.758	2	0.45

Table 6-2. Continued

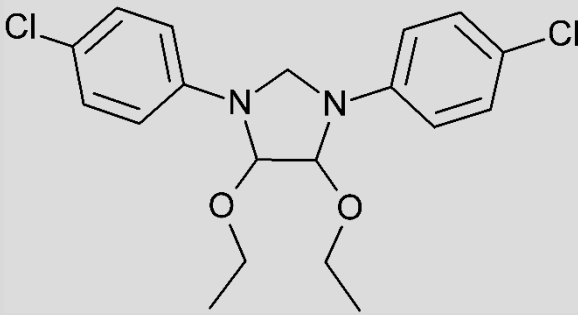
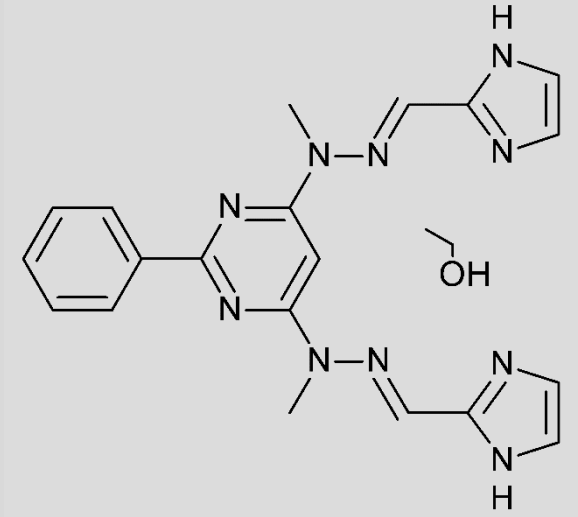
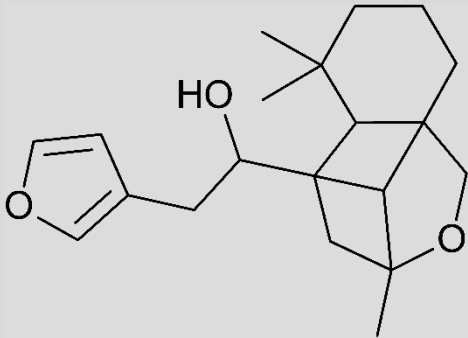
IGASOZ ²² 	3.76	0	0.812
SEHSOP ²³ 	3.76	2	0.449
From methanol data			
AFETIO ²⁴ 	22	0	0.686

Table 6-2. Continued

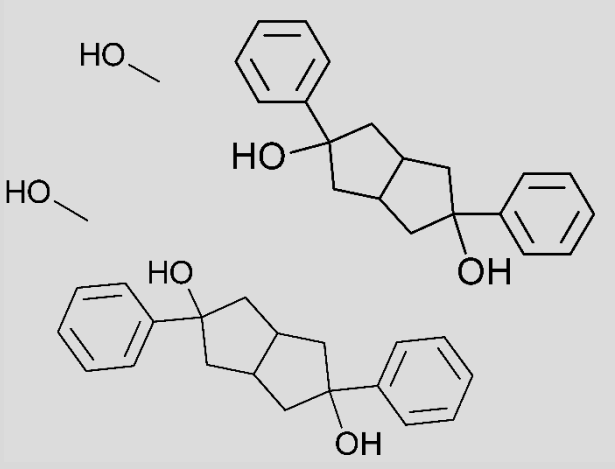
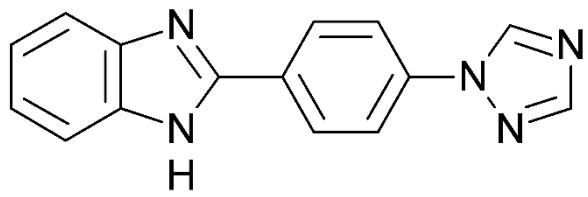
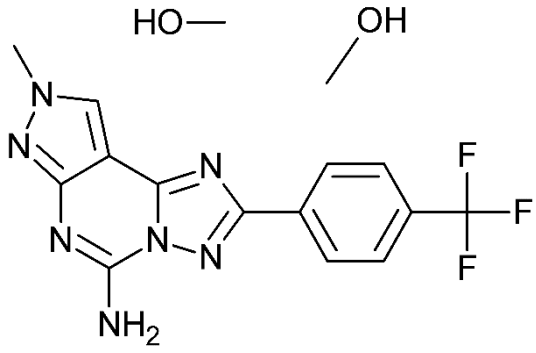
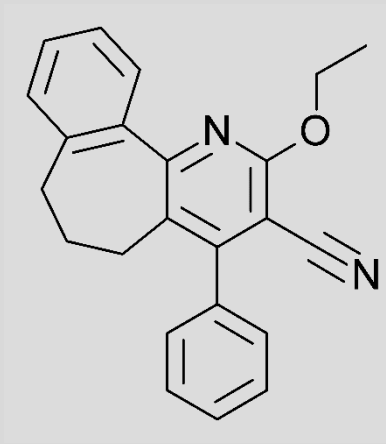
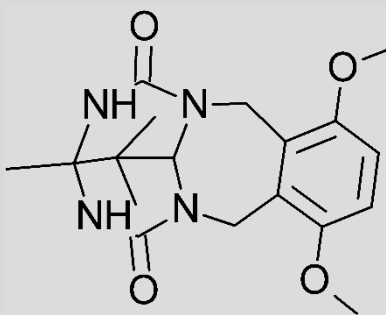
<p>ANOQOK²⁵</p> 	22	2	0.391
<p>CEHCOJ²⁶</p> 	22	1	0.542
<p>YUQKAX²⁷</p> 	22	2	0.391

Table 6-2. Continued

<p>OJOHUQ²⁸</p> 	25	0	0.629
<p>QUCSEM²⁹</p> 	25	2	0.332

6.2 Effects the models do not take into account

The previous section has shown examples to prove the usefulness of the descriptors that are included in the models. Bearing in mind that 20 % of the data was not correctly predicted by the models, there must be factors that account for solvate formation which the models did not consider. These factors are going to be shown by examples from the datasets.

6.2.1 Hydrogen bond strengths

The addition of hydrogen bonding to the models has increased their prediction ability, especially in the models of the alcohol containing solvents. Nevertheless, the description of this property was too simple. The count of the hydrogen bond donors is certainly not the best description of the hydrogen bonding ability of a given molecule. This is due to the fact that the strength of hydrogen bonds depends on many factors, such as the nature of the hydrogen bond donating atom. For instance, a primary amine has a lower affinity than an amide to donate a hydrogen bond.³⁰

An example that can demonstrate this difference in ability to donate hydrogen bonds can be seen by comparing 7-hydroxy-1-methyl-N-(9-methyl-9-azabicyclo[3.3.1]non-3-yl)-1H-indazole-3-carboxamide methanol solvate, CSD reference code: VUQMEA³¹ and 4-(4-Fluorophenyl)-1-phenyl-3-(pyridin-4-yl)-1H-pyrazol-5-amine, CSD reference code: LANRUP.³² Both of these entries were recrystallized from methanol. Their structures are shown in Figure 6-5.

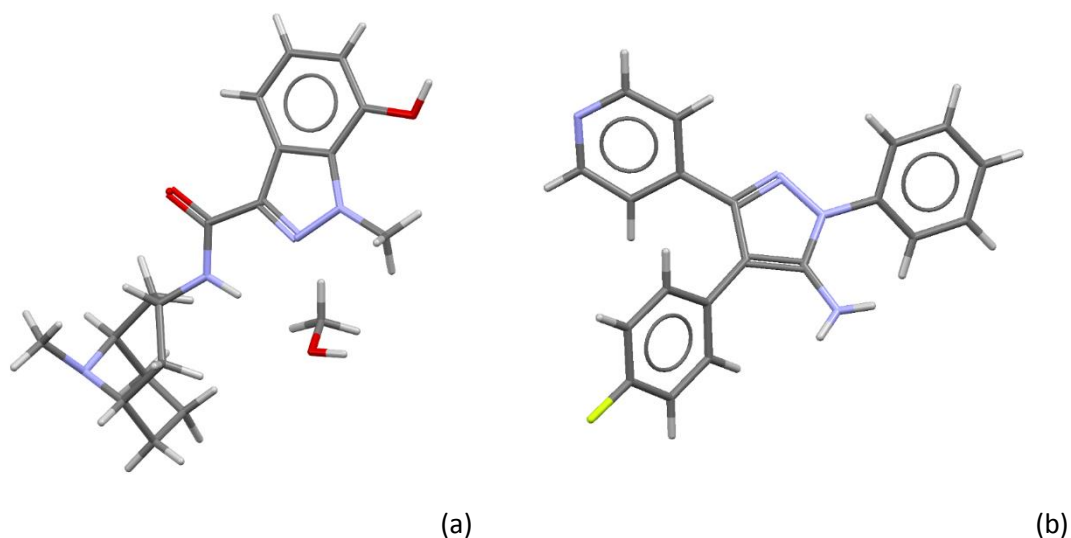


Figure 6-5. Molecular structures of (a) VUQMEA and (b) LANRUP.

The values of the descriptors TRS and nHDon are identical for both compounds; these are 23 and 2, respectively. While the count of hydrogen bond donating groups is the same, the nature of these groups is not. The VUQMEA entry shows an amide and a hydroxyl group while the LANRUP entry has a primary amine. Amides are known to form stronger hydrogen bonds than primary amines; therefore the former is expected to form a stronger interaction than the latter.³³ In the VUQMEA entry, the hydrogen bond between methanol and the amide group, in which the (N–H...O distance is below 2.5 Å) is thought to cause the retention of the solvent in the crystal. Despite this difference in hydrogen bond donating ability, the methanol predictive model looks at them as identical structures. It predicts both of them to form a solvate at value of $x=0.373$. In fact, it correctly predicts the solvate form of the VUQMEA entry but fails to predict the behaviour of the LANRUP entry.

Another example is 2-((2-Aminophenylimino)(phenyl)methyl)-4-chlorophenol, CSD reference code: DEMFUX³⁴ and N'-(2-hydroxynaphthylidene)-3-hydroxybenzohydrazide methanol solvate, CSD reference code: LOMLEF.³⁵ These 2 entries are illustrated in Figure 6-6.

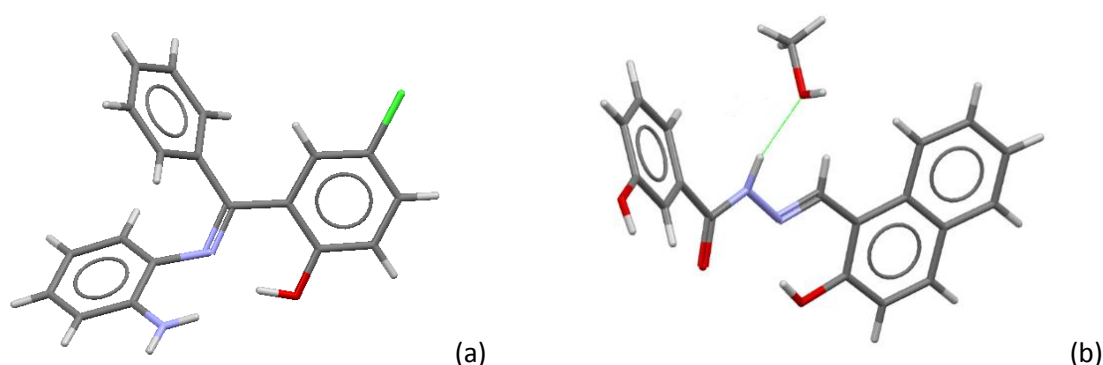


Figure 6-6. An illustration of the DEMFUX (a) and the LOMLEF (b) entries.

From the methanol model perspective, these two entries are identical as they share the same values for the descriptors in that model. They both possess a TRS value of 18 and an nHDon value of 3. Consequently, they both were predicted to form a solvate at a value of 0.330. By

looking at the illustrations of the molecules in Figure 6-6, it can be seen that the hydrogen bond donor groups are quite different. The DEMFUX entry shows one primary amine and a hydroxyl group while the LOMLEF entry shows two hydroxyl and one amide group. The misprediction of the DEMFUX is thought to be caused by the inability of the current models to differentiate between a strong and a weak hydrogen bonding group.

6.2.2 Accessibility of hydrogen bonding

Being one of the main factors involved in the formation of the solvate form, the hydrogen bonding ability of the misclassified compounds was investigated more closely. The nature of the hydrogen bond donor is not the only factor that influences this type of interaction. It can be influenced by other factors, such as the availability of the hydrogen bond donors. In many instances, molecules possessing hydrogen bond donor groups have failed to make hydrogen bonds with the solvent due to their inaccessibility or involvement in intramolecular bonding. An example of that can be seen in the case of tris(2-hydroxy-3-*t*-butyl-5-methoxybenzyl)amine, CSD reference code: RACMEP³⁶ from the methanol dataset. The 3D structure is illustrated in Figure 6-7.

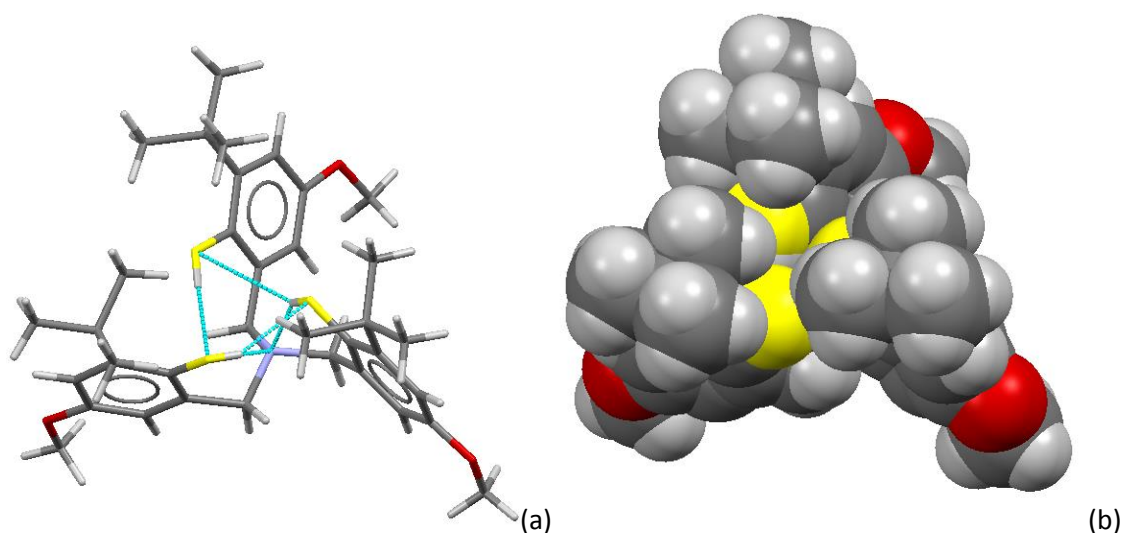


Figure 6-7. An illustration of the limited accessibility of the hydrogen bond donors (the oxygen atoms in yellow) in the RACMEP entry. Part (a) of the Figure shows the capped sticks model of the entry; part (b) shows the space-filling model of the same entry. The intramolecular hydrogen bonds are shown in the blue lines between atoms in part (a).

This molecule has failed to form a solvate upon crystallization from methanol. It has a TRS value of 18 and possesses 3 hydrogen bond donors. As the capped-stick model in **Figure 6-7(a)** shows, the three hydrogen bond donating groups are involved in intramolecular hydrogen bonds. Additionally, the space-filling model in **Figure 6-7(b)** shows that these donors are hardly accessible for the solvent molecules due to the steric effect, leaving them with a low accessible surface area. For these reasons, the methanol model has failed to correctly predict the solvate formation of this entry, where it gave an x value of 0.330.

Another example from the methanol dataset is the case of 5,11,17,23-tetra-*t*-butyl-25,27-*bis*-(2-(*N*-(pyrid-3-ylcarbonyl)amino)ethoxy)-26,28-dihydroxycalix[4]arene, CSD reference code: AZOMIL.³⁷ The entry is illustrated in **Figure 6-8**.

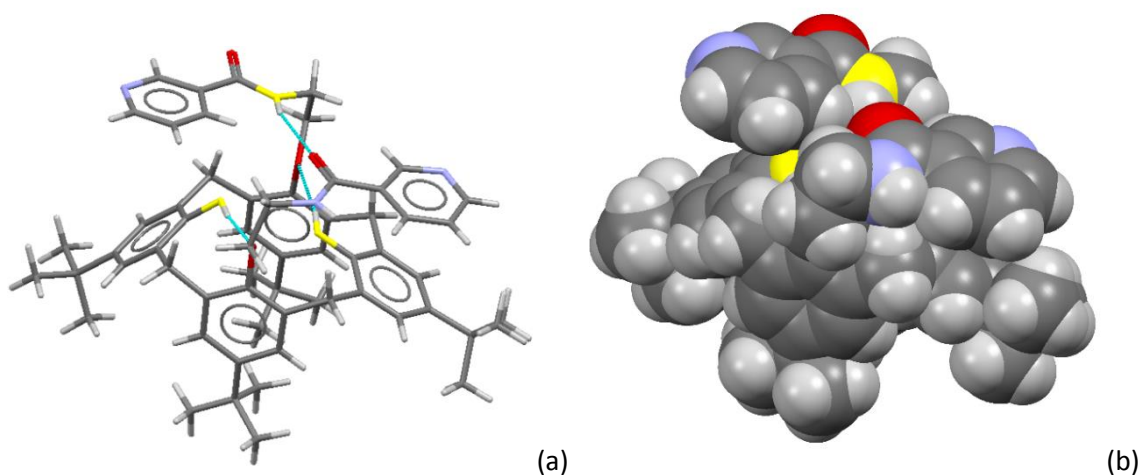


Figure 6-8. (a) Capped-stick and (b) space filling representation of the AZOMIL molecule. The inaccessible donors (oxygen atoms) are highlighted in yellow. The intramolecular hydrogen bonds are shown in the blue lines between atoms in part (a).

Similar to the previous example, this entry has shown a high TRS (52) value and a large number of hydrogen bond donors ($n\text{HDon}=4$). This compound was recrystallized from methanol and has failed to form a solvate. The likely reason for the inability of this entry to form a solvate is the low availability of the hydrogen bond donors. Three of the four donor groups in this structure participate in intramolecular hydrogen bonds. This can be seen by looking at Figure 6-8(a) where the capped stick model clearly shows it. The other factor is the steric effect where other atoms have prevented the solvent from interacting with the donors. This can be noticed by looking at Figure 6-8(b) which shows the low accessible surface area. The methanol model has failed to correctly predict the solvate formation of this entry, where it gave an x value of 0.015. This large misprediction is mostly attributed to the reasons mentioned. A large number of examples exists in the datasets. The reference codes of these entries are given in

Table 6-3. Similar to Table 6-2, this table shows examples from ethanol and methanol datasets because their models were the only two improved by the addition of the hydrogen bond factor to the models.

Table 6-3. Accessibility of hydrogen bond donors

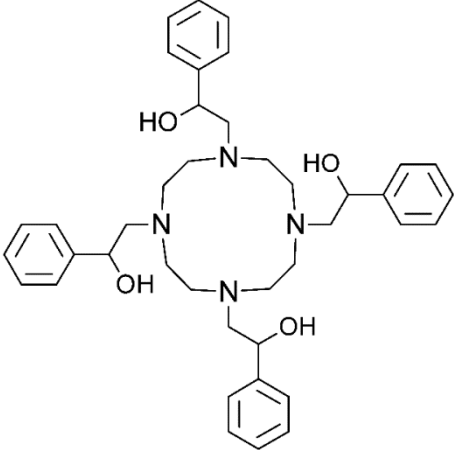
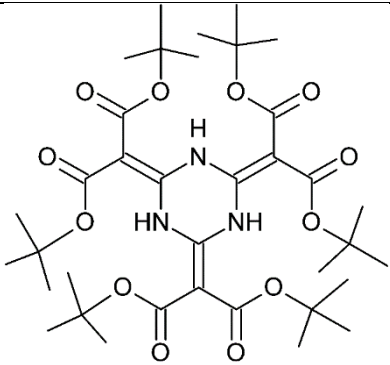
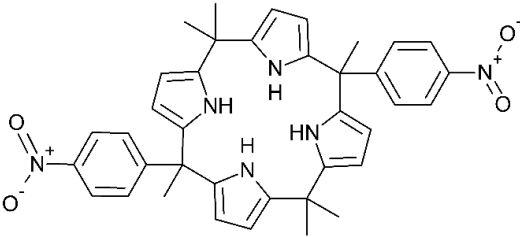
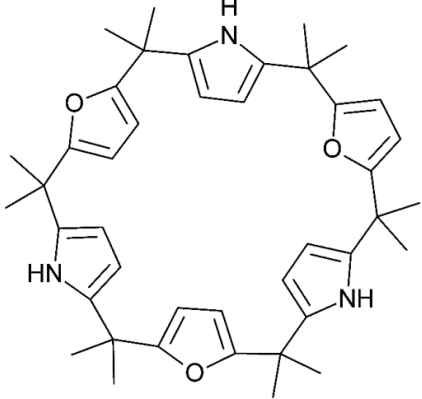
From ethanol data				
CSD reference code	AVS_H2	nHDon	Chemical formula	Prediction value
ICOTIE ³⁸	3.938	4		0.069
QEXYUO ³⁹	4.04	3		0.106
GILZOR ⁴⁰	4.474	4		0.009
WISTUN ⁴¹	4.455	3		0.024

Table 6-3. Continued

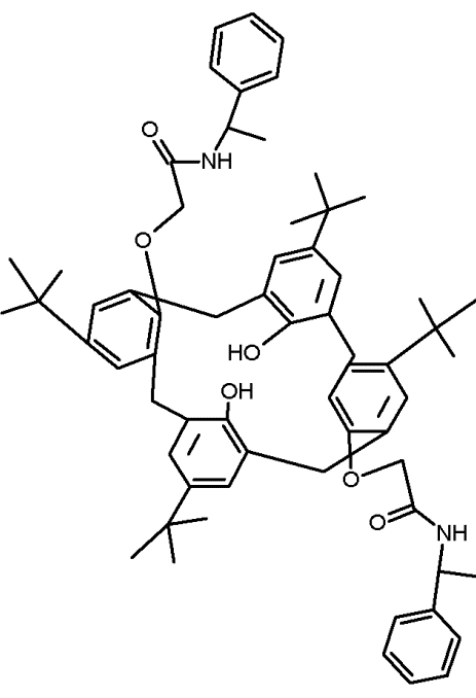
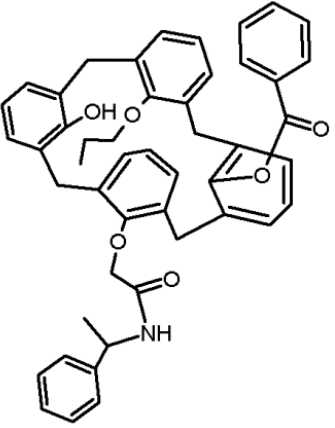
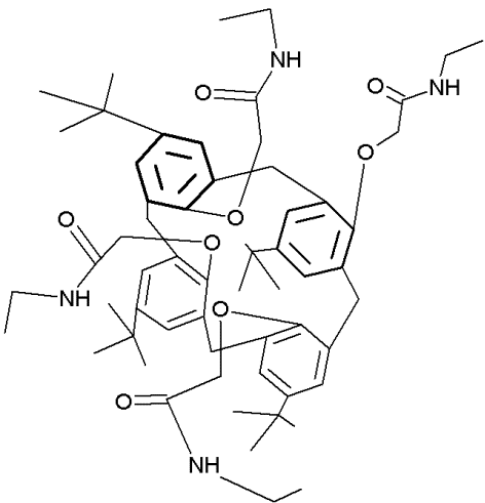
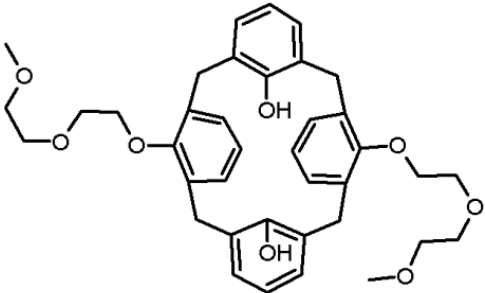
From methanol data				
CSD reference code	TRS	nHDon	Chemical formula	Prediction value
AWITIJ ⁴²	52	4		0.015
TUFSIX ⁴³	52	2		0.049

Table 6-3. Continued

QASFUL ⁴⁴	40	4		0.040
RUWGOF ⁴⁵	40	2		0.123

6.2.3 Ring interactions

This type of interaction was widely observed throughout the dataset at short contact distances (see the beginning of chapter 6 for the definition of short distances). In order to know if this interaction affects solvate formation, a comparison between the short interactions in the solvate and the non-solvate groups in each solvent can be conducted. Note that the search should cover all three common conformations (parallel, offset parallel and T-shaped) of the ring interactions. In addition to the distance, the angle of the interaction and how much it deviates from the ideal interaction (90 ° for T-shaped, 0 ° for offset and π - π stacking) should be taken into account. An illustration of the deviation from a perfect interaction is illustrated in Figure 6-9.

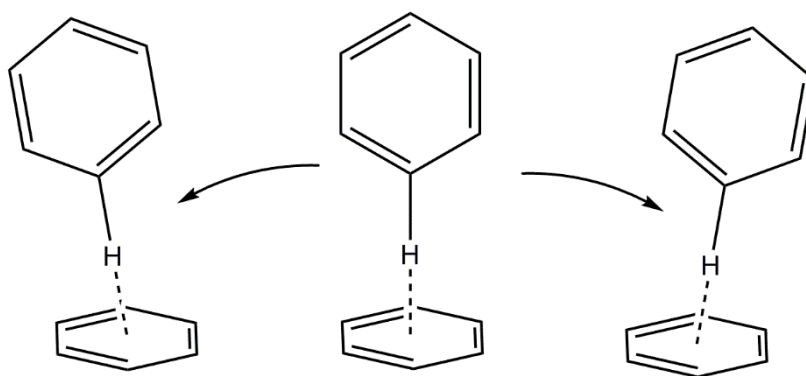


Figure 6-9. An illustration of the deviation that was allowed for the T-shaped interaction from 90 °.

In order to optimize the angle and make it consistent when comparing the distance of interactions, the entries of all solvent datasets (~19,000 molecules) were searched for any type of ring interaction in the range from 3.2 to 5 Å. This range covers the distance from 0.5 Å below to 0.1 Å over the equilibrium distance of the benzene dimer interactions (see section 2.4.3 for more details). The results of this search are shown via a heat map in Figure 6-10.

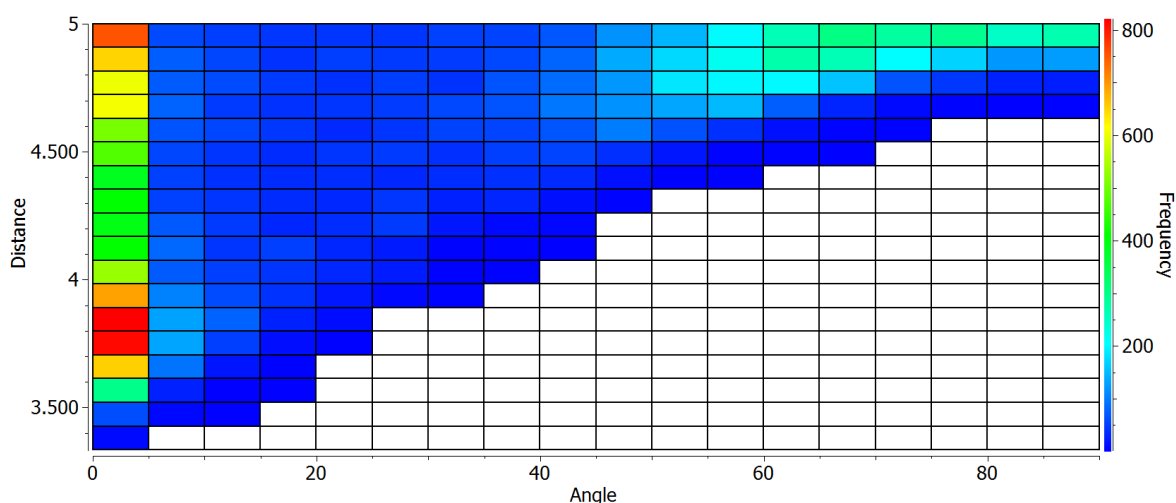


Figure 6-10. A heat map showing the number of entries that from a π - π interaction, where the x-axis shows the angle and the y-axis shows the centroid to centroid distance.

Three regions of high density can be noticed in Figure 6-10, two to the left and one to the right of the figure. The one to the left high density in the range of 3.5 to 4 Å represents the

interactions at an angle near zero. This suggests that entries possessing offset π - π interactions and π - π stacking are in this region. It could be noticed that the interaction is restricted to being completely parallel, where a deviation of 5 ° from the perfectly parallel interaction resulted in almost no hits. The high density region above that, near 5 Å is not thought to represent any binding interactions, as this distance is well above the interaction distance for parallel π - π interactions. The region to the right shows the largest number of interactions around 65 ° indicating the entries on this side are the T-shaped (edge-to-face) interactions. More flexibility can be noticed in this type of interaction, where the first column with completely navy blue boxes can be seen around 40-50 °. As a result, the search that was conducted to count the number of interactions in each solvent allowed for a magnitude of deviation of 10 ° for the parallel interactions (sandwich (parallel) and the offset-parallel interactions) from 0 °, while a deviation of 45 ° from the 90 ° ideal interaction was allowed for the T-shaped interactions. The results of this search are shown in Figure 6-11. The distance on the other hand was restricted to the same concept of short interactions defined in at the beginning of chapter 6.

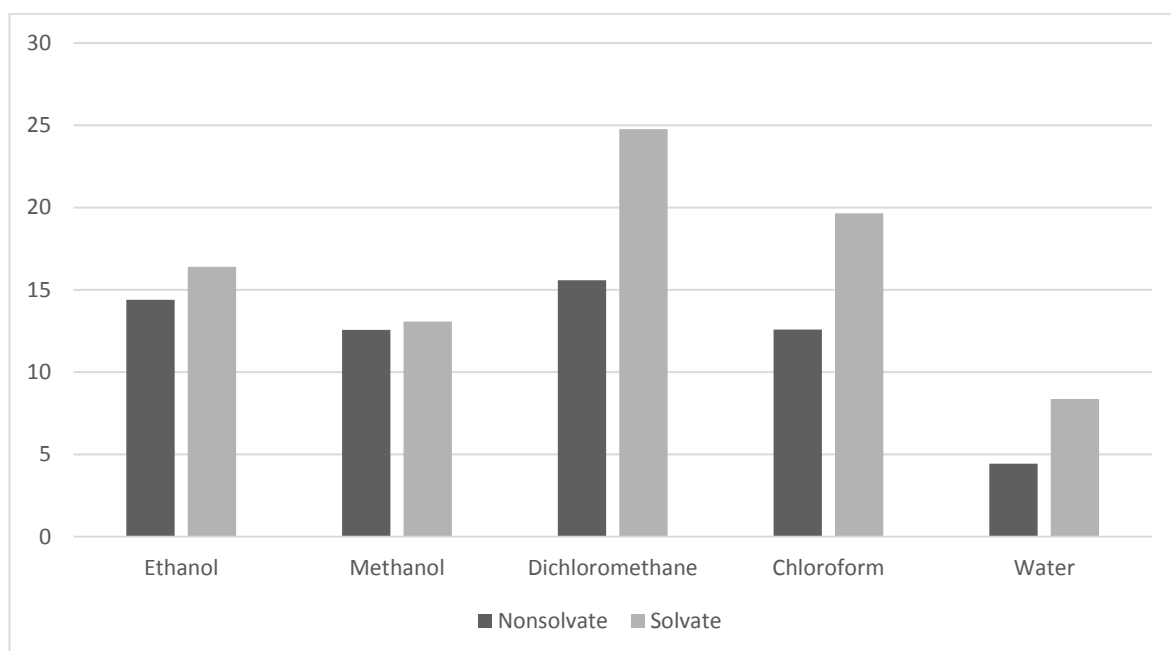


Figure 6-11. The percentage of structures with short ring interactions in the solvate and the non-solvate groups per solvent. Note that these hits are the ones that followed the restrictions set on the angles of the interaction.

Strong ring interactions turned out to be more frequent in solvates. This indicates that ring interactions could aid hydrate and solvate formation to different extents among solvents, with the chlorinated solvents and water being the most affected type of solvents.

6.2.4 Halogen bonding

The chlorinated solvents tested, dichloromethane and chloroform, have drawn attention to the halogen bonding role in solvate formation. More than 25 % of the chloroform solvate entries have shown halogen bond at short distances between the solvent and the molecules (see the beginning of chapter 6 for short distance definition). This could be an evidence of the participation of this type of bonding in solvate formation. On the other hand, no entries in this dataset showed to have a solvate where the solvent is stabilized through halogen bonding alone, which means the chlorinated solvent in any observed crystal always had other interactions binding it. This could suggest that halogen bonding is not strong enough to stabilize a solvate structure on its own.

An example of a positive contribution of this bond towards solvate formation can be seen in *t*-butyl (1-((4-bromophenyl)sulfonyl)-4-(4-methyl-1H-1,2,3-triazol-1-yl)piperidin-3-yl)carbamate chloroform solvate, CSD reference code: KUWWOP.⁴⁶ An illustration is shown in Figure 6-12.

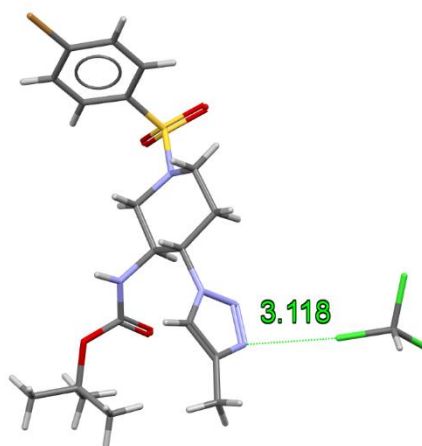


Figure 6-12. Molecular structure of KUWWOP.

This entry shows a chlorine atom at a short distance (3.118 Å) from a nitrogen atom (C-Cl...N angle=173.8 °). It has a SM3_H2 value of 4.663 and an H.050 value of 1, both of which led to a prediction value of $x = 0.520$ by the chloroform model (formation of non-solvates). The prediction of this chloroform solvate was incorrect where the probability found was 0.020 over the cut-off point of the chloroform model. The inclusion of halogen bonding in the model could probably shift the predictions of such a case towards the correct prediction region.

Another example from the chloroform dataset is *meso*-1,12-dimethylene-2,11-dithia[3.3]metacyclophane 2,11-dioxide chloroform solvate, CSD reference code: PIGDOZ.⁴⁷ The 3D structure of the entry is shown in Figure 6-13.

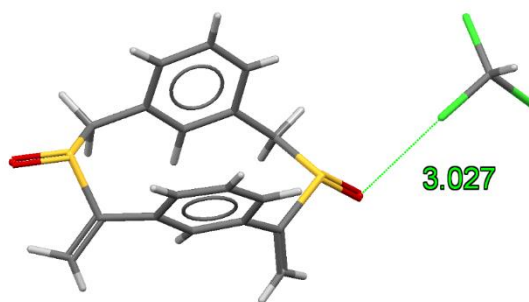


Figure 6-13. An illustration of the halogen bond in the PIGDOZ entry. It shows the role of this type of bonding in the chloroform solvate formation. The bond takes place at a distance 0.1 Å shorter than the sum of the van der Waals radii of the two atoms.

This entry shows a chlorine-oxygen halogen bond at a short distance of 3.027 Å (C-Cl...N angle=178.06 °). The molecule has an SM3_H2 value of 4.339 and a H.050 value of 0. It was predicted not to form a solvate at x=0.810. A number of entries in different datasets have also shown the effect of this type of interaction. Additional examples from the chloroform dataset are given in Table 6-4.

Table 6-4. Additional examples of mispredicted data, involving a short halogen bond in the structure. The examples are given from the chloroform dataset

From dichloromethane data				
CSD reference code	SM3_H2	Hy	Chemical formula	Prediction value
SUQQEA ⁴⁸	4.560	-0.601		0.656

Table 6-4. Continued

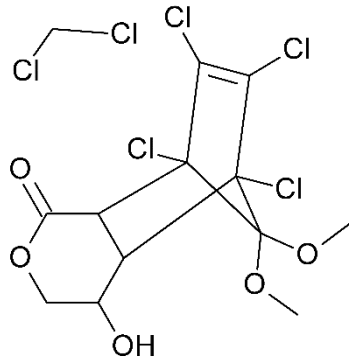
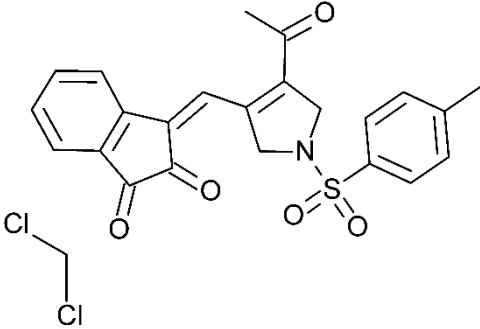
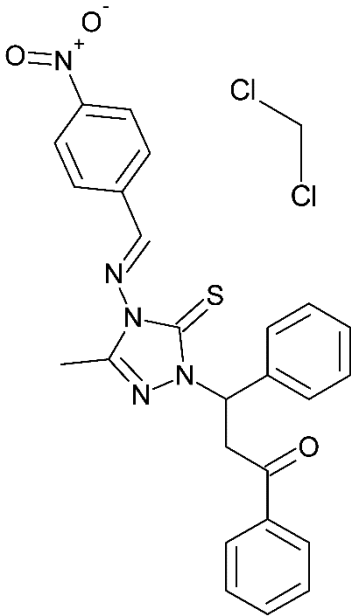
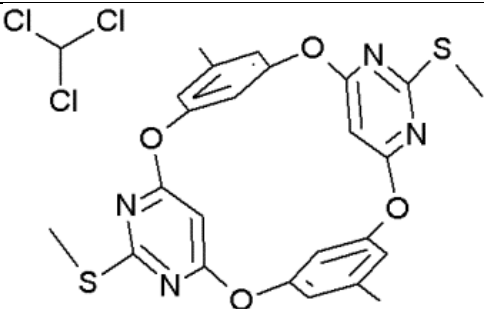
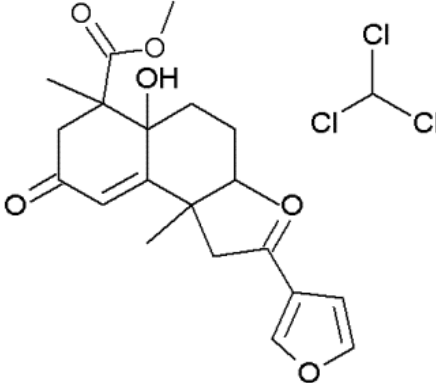
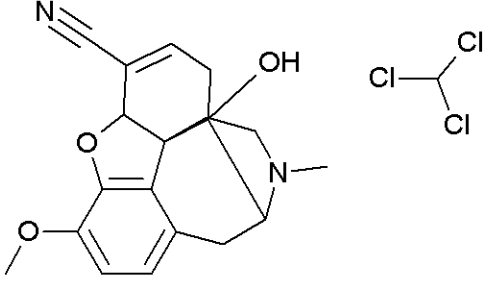
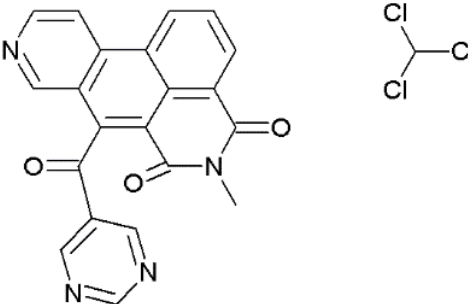
VOHHAC ⁴⁹	4.603	-0.104		0.543
MOHTOS ⁵⁰	4.741	-0.759		0.537
YAJCOD	4.742	-0.729		0.531

Table 6-4. Continued

From chloroform data				
CSD reference code	SM3_H2	H.050	Chemical formula	Prediction value
REMDUK ⁵¹	4.805	0		0.508
FIJXOM ⁵²	4.686	1		0.503
PASMUS ⁵³	4.765	0		0.539
KAJRU ⁵⁴	4.747	0		0.552

6.2.5 Zwitterions

Some entries were mispredicted despite the fact that they had a prediction with a high confidence. Examples of entries that were predicted with a high confidence for solvate formation but failed to form them were given in section 6.2.2. This has happened when the hydrogen bond donors were not accessible by the solvent molecules, leading to an unexpected non-solvate. The opposite case was when some entries having a small size, low branching and small number of hydrogen bond donors were able to form a solvate. A closer look on these entries have revealed the fact that many of them were zwitterionic, i.e. having a formal positive and a formal negative charge on the same molecule. These entries were noticed when the data was first investigated in chapter 4 (section 4.3.2) but were not excluded from the datasets. An example representing this case is 1-ammoniocyclopropanecarboxylate hemihydrate, CSD reference code: FOBJUB55 (Figure 6-14).

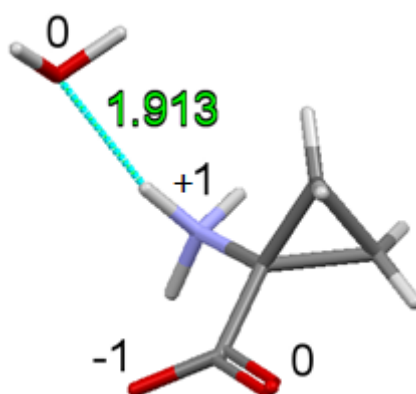


Figure 6-14. Molecular structure of FOBJUB. The heteroatoms are labelled with their formal charge.

This entry has an nH value of 7 and a π ID value of 3.871. These values have led to a probability of $x=0.905$ by the hydrate model, which is far from reality. The entry shows an unusually short interaction between a hydrogen bond donating group (ammonium ion (NH_3^+)) and the oxygen of water at a distance of 1.913 Å. The large misprediction associated with this entry is mostly related to donor strength where the charged nitrogen is involved. The formal positive charge

might have caused it to be more electronegative than usual, leading to a strong interaction with the water molecule. This was not the only case witnessed in the datasets. The role of this highly polar nitrogen atom in hydrate formation coincides with what has been reported by L. Infantes *et al.* in a CSD investigation of around 35,000 organic molecules in 2006.⁵⁶ Plenty of examples on this type of interaction were seen in each solvent dataset, which gives a warning sign to the scientists interested in developing charged organic materials. A list of examples illustrating this case from the water data are given in Table 6-5.

Table 6-5. Additional examples on mispredictions due to the presence of a zwitterion

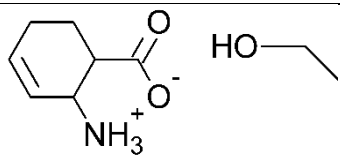
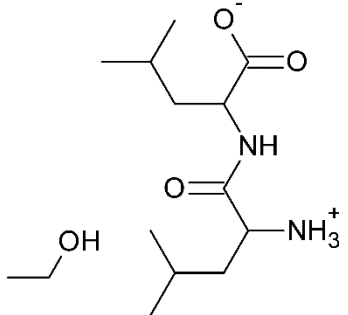
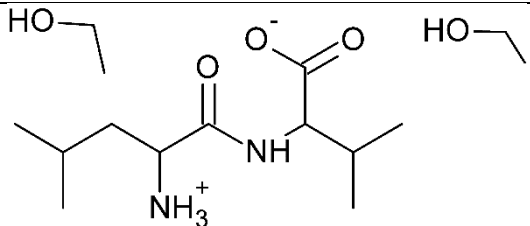
From ethanol data				
CSD reference code	AVS_H2	nHDon	Chemical formula	Prediction value
APUNAA ⁵⁷	3.033	3		0.847
JUQQIV ⁵⁸	3.107	4		0.638
SUWLLOL ⁵⁹	3.097	4		0.646

Table 6-5. Continued

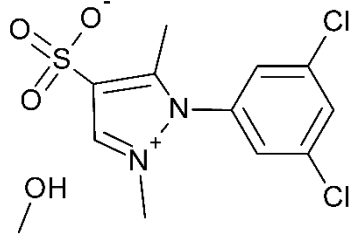
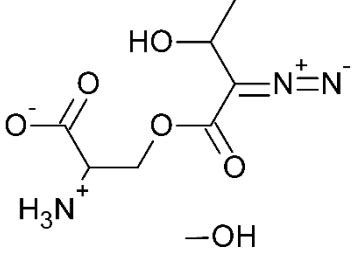
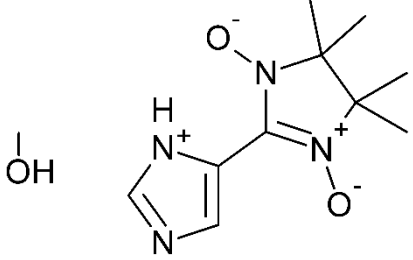
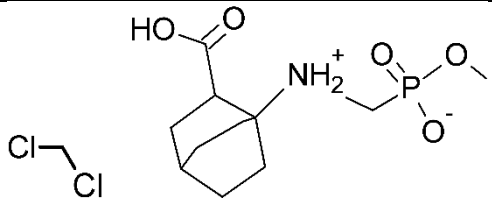
From methanol data				
CSD reference code	TRS	nHDon	Chemical formula	Prediction value
FEMQEV ⁶⁰	11	0		0.848
SAWGOM ⁶¹	0	4		0.551
AKAWUE ⁶²	10	1		0.767
From dichloromethane data				
CSD reference code	SM3_H2	Hy	Chemical formula	Prediction value
EVODAV ⁶³	4.151	1.306		0.676

Table 6-5. Continued

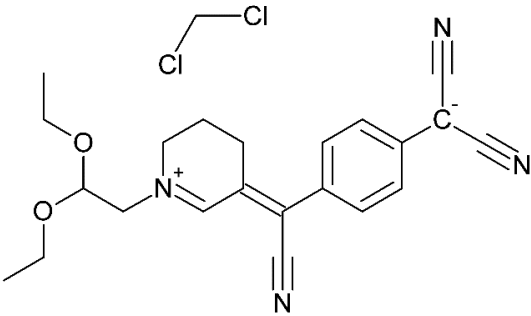
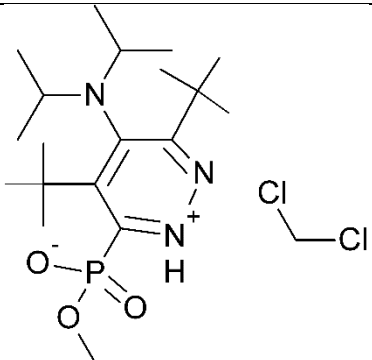
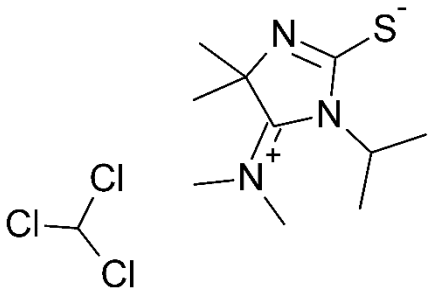
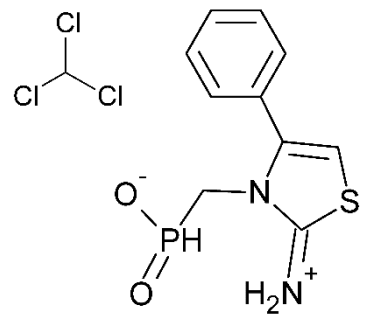
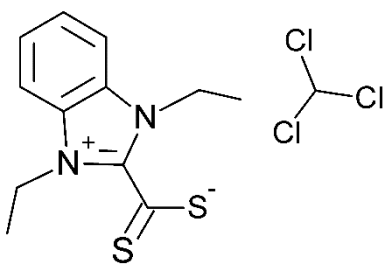
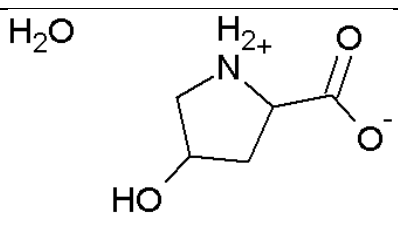
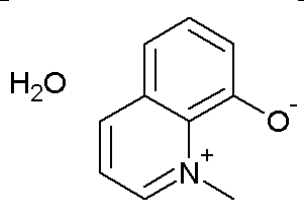
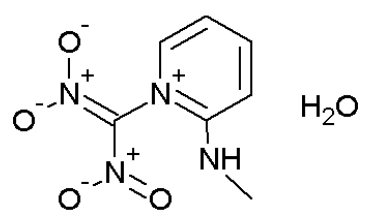
ROLDEB ⁶⁴	4.420	-0.778		0.773
DAHME ⁶⁵	4.569	-0.294		0.601
From chloroform data				
CSD reference code	SM3_H2	H.050	Chemical formula	Prediction value
MIPIMT ⁶⁶	3.886	0		0.945
TEQTUG ⁶⁷	3.926	3		0.827

Table 6-5. Continued

WAQLAD ⁶⁸	4.019	0		0.919
From water data				
CSD reference code	π ID	nH	Chemical formula	Prediction value
AHLPRO ⁶⁹	4.477	9		0.860
CEDJEA ⁷⁰	7.833	9		0.588
MAPYNM ⁷¹	7.008	8		0.691

6.2.6 Why these factors were not included in the predictive models

Multiple factors that the models did not account for were mentioned in sections 6.2.1 to 6.2.5.

A sensible question here would be why the statistical investigation did not identify any of these factors. The answer to this question can be one of these cases:

- The number of examples of a specific factor might not have been large enough to be statistically significant. An example of this case can be the zwitterions problem (section 6.2.5), where about 30 examples representing the case were found in each dataset. The fact that the smallest dataset is made of more than 2,500 data points, made these entries insignificant.
- The factor is partially encoded in a descriptor that is already in the model. For example, the ring interactions (section 6.2.3) is a factor that was highly involved in the solvate formation of the chlorinated solvents. Nevertheless, it was not one of the three top descriptors that showed up in the statistical analysis, not even as the third most important descriptor. This is because the descriptors found by the logistic regression model (spectral moment in the case of chloroform and dichloromethane) inherently include information about the number of rings. For example, the correlation between the spectral moment descriptor SM3_H2 and the number of rings nCIC in the chloroform dataset is larger than 0.75.
- No descriptor is available to represent that factor. An example of this is the hydrogen bond strength, where the descriptors available give either a yes or no, but do not give a weighing to the hydrogen bond.

6.3 Possible improvements

At this point, it would be valuable to suggest solutions to improve the predictive ability of the models. This section suggests possible improvements that can solve some issues mentioned in section 6.2, leading to an effective, simple prediction of the solvate formation.

6.3.1 Inclusion of a hydrogen bond strength scale

The hydrogen bond donating ability can vary depending on the donating group and the environment surrounding it. Section 6.2.1 has explained this variability with examples. The question remains on how could the description of the hydrogen bond donating ability be improved? To start with, researchers are aware of this variation as mentioned in several publications.^{72, 73} Different research groups have worked on finding a solution to quantify this interaction. In 2001, a study led by Michael H. Abraham from the University College London presented empirical hydrogen bond structural constants indicating the relative ability of functional groups to donate and accept hydrogen bonds. These constants were obtained based on the equilibrium constant of a 1:1 hydrogen bonded complex. They were provided for both aromatic and aliphatic functional groups. In 2004, another group, led by Christopher A. Hunter was able to provide similar description based on the electrostatic potential surface of molecules. Despite the different approaches taken by these two groups, the ranking of hydrogen bonding ability between functional groups is analogous.

Incorporating these findings into the current predictive models can improve their predictive ability. To be precise, it could be most beneficial to the predictions that are close to the decision boundary of the models, where a small shift towards the correct prediction region can influence the outcome. This assumption can be demonstrated by referring to the example in Section 6.2.1. The first entry in that example (CSD reference code: VUQMEA), which formed a methanol solvate possessed an amide group. The second entry (CSD reference code: LANRUP)

which has failed to form a solvate possessed a primary amine group. The empirical scales set by Abraham suggests that the hydrogen bond donating constant for the aliphatic amide group ranges between 0.40-0.55 compared to 0.08-0.16 for aliphatic amines. The amine in the LANRUP structure is connected to a benzene ring, which might mean it is close to being an aromatic amine due to the delocalization of the ring. The hydrogen bond donating constant for aromatic amines range is 0.17-0.26, which is still significantly lower than the constant of the amide.

Although the hydrogen bond acceptors are not part of the predictive models at this point, an illustrative example is essential to prove the usability of these scales. The comparison between (3*S*,6*S*)-3,6-Diethyl-3,6-dimethyl-1,4-*bis*(1-phenylethyl)piperazine-2,5-dione methanol solvate, CSD reference code: IXAYOW⁷⁴ and *cis*-1,4-dicyano-7,7-dimethoxy-6,6-dimethyl-2,3-benzo-*cis*-bicyclo[3.3.1]nonane, CSD reference code: QIHFAO⁷⁵ from the methanol dataset, suggests the importance of the hydrogen bond accepting scales. Both of these structures are shown in Figure 6-15.

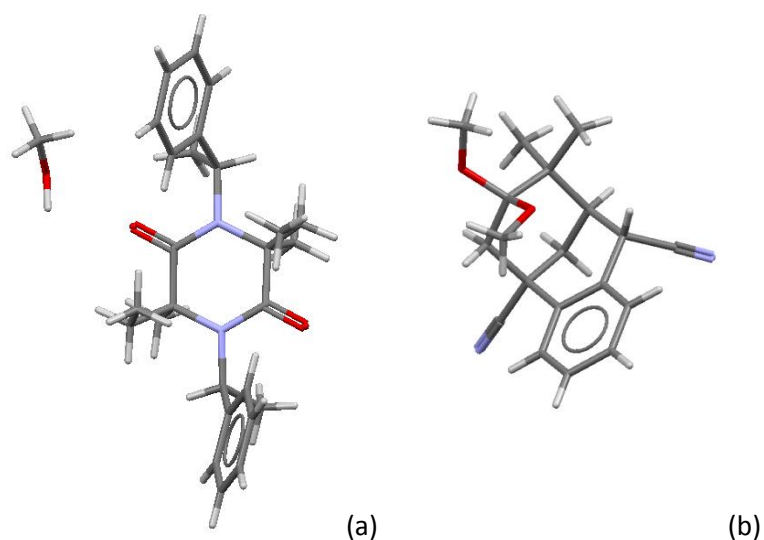


Figure 6-15. Molecular structures of IXAYOW and QIHFAO.

These two molecules possess hydrogen bond acceptors. These groups are available for interaction with methanol. They both had a TRS descriptor value of 18 and an nHDon descriptor value of 0. The model does not qualify them to form a methanol solvate, where they are predicted at a value of $x = 0.755$. This means that the model correctly predicts the QIHFAO entry but fails to predict IXAYOW. The reason for this misprediction is thought to be due to the difference in the hydrogen bond accepting ability. In the paper published by Abraham, it is shown that the carbonyl of the amide group acts as a stronger acceptor than the ether or cyanide groups. The assigned values for their strengths on the hydrogen bond accepting scale is 0.77-0.80 for amide while it ranges between 0.48-0.51 and 0.36-0.44 for ether and cyanide groups, respectively.⁷²

6.3.2 Application of Etter's rules

One of the issues that limited the ability of the models to correctly predict the solvate formation was the inaccessibility of the strongly-binding functional groups. One reason for this inaccessibility was intramolecular hydrogen bonding or involvement of hydrogen bond donor/acceptor groups in intramolecular hydrogen bonding (see section 6.2.2). Margaret C. Etter has proposed a list of rules regarding hydrogen bonding.⁷⁶ These rules focus mainly on the preferred interactions for strong hydrogen bond donors and acceptors. For example, one of the general rules states that *"The best proton donors and acceptors remaining after intramolecular hydrogen-bond formation form intermolecular hydrogen bonds to one another."* The value of this rule can be seen in the cases illustrated and tabulated in section 6.2.2. The incorporation of these rules into the models has the potential to improve their predictive ability.

6.3.3 Inclusion of steric hindrance description

The application of Etter's rules can certainly account for a considerable part of the mispredictions associated with inaccessibility of the functional groups. Another reason why some groups were inaccessible was the steric hindrance around these groups. But how can this be studied theoretically? A model of hydrogen bond propensity in organic crystals, known as the "logit hydrogen bond propensity" (LHP) was introduced by P. Galek *et al.* in 2007.⁷⁷ One of the parameters used by this model was called the steric density function. This function describes the density of non-hydrogen bonding atoms in the environment around the hydrogen bonding functional groups. When there is a high density of non-hydrogen bonding atoms, this means the hydrogen bond donor/acceptor is less likely to be accessible by other atoms. Such a function could be implemented as part of the predictive models. In more technical terms, the knowledge of the chemical structure of a molecule, along with the knowledge of the average bond lengths and atomic sizes can help find the density of atoms in a specific area, leading to an anticipation of the availability of the hydrogen bond donating/accepting groups.

6.4 References

1. Garbelini ER, Hörner M, Giglio VF, da Silva AH, Barison A, Nunes FS. Synthesis, Crystal Structure and Spectroscopic Properties of Bis(1-(E)-2-formylpyridine semicarbazone)nickel(II) Diperchlorate Monohydrate. *Zeitschrift für Anorganische und Allgemeine Chemie*. 2009;635(8):1236-41.
2. Rapheal PF, Manoj E, Kurup MRP, Suresh E. N-(Pyridin-2-yl)hydrazinecarbothioamide. *Acta Crystallographica Section E*. 2005;61(7):o2243-o5.
3. Décor A, Monse B, Martin M-T, Chiaroni A, Thoret S, Guénard D, et al. Synthesis and biological evaluation of B-ring analogues of (–)-rhazinilam. *Bioorganic & Medicinal Chemistry*. 2006;14(7):2314-32.
4. Cruz S, Trilleras J, Cobo J, Low JN, Glidewell C. Four N7-benzyl-substituted 4,5,6,7-tetrahydro-1H-pyrazolo[3,4-b]pyridine-5-spiro-1'-cyclohexane-2',6'-diones as ethanol hemisolvates: similar molecular constitutions but different crystal structures. *Acta Crystallographica Section C*. 2008;64(12):o637-o42.
5. Middel O, Greff Z, Taylor NJ, Verboom W, Reinhoudt DN, Snieckus V. The First Lateral Functionalization of Calix[4]arenes by a Homologous Anionic Ortho-Fries Rearrangement. *The Journal of Organic Chemistry*. 2000;65(3):667-75.
6. Shi Y-Y, Sun J, Huang Z-T, Zheng Q-Y. Crystalline Self-Assembly of a Bowl-Like Cyclotriguaiacylene Derivative with Alcohol/Phenols by Hydrogen Bonding and C–H⋯ π Interactions: The Self-Inclusion Extended Organic Frameworks. *Crystal Growth & Design*. 2010;10(1):314-20.

7. Lee HN, Xu Z, Kim SK, Swamy KMK, Kim Y, Kim S-J, et al. Pyrophosphate-Selective Fluorescent Chemosensor at Physiological pH: Formation of a Unique Excimer upon Addition of Pyrophosphate. *Journal of the American Chemical Society*. 2007;129(13):3828-9.
8. Steiner T, Tamm M, Grzegorzewski A, Schulte N, Veldman N, Schreurs AMM, et al. Weak hydrogen bonding. Part 5. Experimental evidence for the long-range nature of $C\equiv C-H\cdots\pi$ interactions: crystallographic and spectroscopic studies of three terminal alkynes. *Journal of the Chemical Society, Perkin Transactions 2*. 1996(11):2441-6.
9. Scott JL, Parkin SR, Anthony JE. Radical-Induced Cycloaromatization: Routes to Fluoranthenes and Acephenanthrylenes. *Synlett*. 2004;2004(1):161-4.
10. Koshima H, Nakata A, Nagano M, Yu H. Photoreaction of 9-Methylbenz [c] acridine with Diphenylacetic Acid in the Chiral Cocrystal. *Heterocycles*. 2003;60(10):2251-8.
11. Wang X-R, Xing J, Yan C-X, Cheng Y. The reaction of β -lactam carbenes with 3,6-dipyridyltetrazines: switch of reaction pathways by 2-pyridyl and 4-pyridyl substituents of tetrazines. *Organic & Biomolecular Chemistry*. 2012;10(5):970-7.
12. Khlebnikov AF, Novikov MS, Petrovskii PP, Stoeckli-Evans H. An Aza Cyclopropylcarbinyl-Homoallyl Radical Rearrangement–Radical Cyclization Cascade. Synthesis of Dibenzoimidazoazepine and Oxazepine Derivatives. *The Journal of Organic Chemistry*. 2011;76(13):5384-91.
13. Shi X, Attygalle AB, Meinwald J, Houck MA, Eisner T. Spirocyclic Defensive Alkaloid from a Coccinellid Beetle. *Tetrahedron*. 1995;51(32):8711-8.

14. Alhadi AA, Saharin SM, Mohd Ali H, Robinson WT, Abdulla MA. N'-(2,4-Dimethoxybenzylidene)-3,4,5-trihydroxybenzohydrazide ethanol solvate. *Acta Crystallographica Section E*. 2009;65(6):o1373.

15. Santos-Contreras RJ, Martinez-Martinez FJ, Mancilla-Margalli NA, Peraza-Campos AL, Morin-Sánchez LM, Garcia-Báez EV, et al. Competition between OH...O and multiple halogen-dipole interactions on the formation of intramolecular three-centred hydrogen bond in 3-acyl coumarins. *CrystEngComm*. 2009;11(7):1451-61.

16. Wu W-N, Shi J-C, Li X-X, Qin B-F. Crystal structure of ethyl 5-((2-hydroxybenzoylaminoimino)methyl)-3,4-dimethyl-1H-pyrrole-2-carboxylate — ethanol (1:2), C₁₇H₁₉N₃O₄ · 2C₂H₅OH. *Zeitschrift für Kristallographie - New Crystal Structures* 2010;225(4):725-6

17. Zhao Y-L, Zhang Q-Z, Chen X, Yu M. (E)-N'-[4-(2-Chlorobenzyloxy)benzylidene]isonicotinohydrazide. *Acta Crystallographica Section E*. 2006;62(11):o4928-9.

18. Rajalakshmi P, Srinivasan N, Krishnakumar RV. 4-Cyano-N-ethylspiro[chromene-2,4'-piperidine]-1'-carboxamide. *Acta Crystallographica Section E*. 2013;69(1):o138.

19. Styngach EP, Malinovskii ST, Bets LP, Vlad LA, Gdanets M, Makaev FZ. Crystal and molecular structure of (1S,2S,4aS, 8aS)-N-(N-allyldiaminomethanethione)-1-(2-hydroxy-2,5,5,8a-tetramethyldecahydronaphthalenyl) acetamide. *Journal of Structural Chemistry* (Translation of *Zhurnal Strukturnoi Khimii*). 2005;46(4):765-9.

20. Perisanu S, Contineanu I, Banciu MD, Liebman JF, Farivar BS, Mullan MA, et al. The enthalpies of formation of two dibenzocyclooctadienones. *Thermochimica Acta*. 2003;400(1–2):109–20.
21. Gordon-Wylie SW, Teplin E, Morris JC, Trombley MI, McCarthy SM, Cleaver WM, et al. Exploring Hydrogen-Bonded Structures: Synthesis and X-ray Crystallographic Screening of a Cisoid Cyclic Dipeptide Mini-Library. *Crystal Growth & Design*. 2004;4(4):789–97.
22. Wan Y, Chen X, Zhang P, Wu H. 1,3-Bis(4-chlorophenyl)-4,5-diethoxyimidazolidine. *Acta Crystallographica Section E*. 2008;64(11):o2158.
23. Stefankiewicz AR, Rogez G, Harrowfield J, Sobolev AN, Madalan A, Huuskonen J, et al. Self-ordering of metallogrid complexes via directed hydrogen-bonding. *Dalton Transactions*. 2012;41(45):13848–55.
24. Tanaka J, Marriott G, Higa T, Higa T. Cacofurans A and B, New Furanoditerpenes from a Marine Sponge. *Journal of Natural Products*. 2001;64(11):1468–70.
25. Chan IYH, Bhadbhade MM, Bishop R. Design of a new inclusion host: 3,7-diphenylbicyclo[3.3.0]octane-endo-3,endo-7-diol. *CrystEngComm*. 2011;13(9):3162–9.
26. Cheng L-H, Zheng Z, Han Z-L, Wu Z-C, Zhou H-P. 2-[4-(1H-1,2,4-Triazol-1-yl)phenyl]-1H-benzimidazole. *Acta Crystallographica Section E*. 2012;68(10):o2890.
27. Dolzhenko AV, Tan GK, Dolzhenko AV, Koh LL, Pastorin G. 8-Methyl-2-[4-(trifluoromethyl)phenyl]-8H-pyrazolo[4,3-e][1,2,4]triazolo[1,5-c]pyrimidin-5-amine methanol disolvate. *Acta Crystallographica Section E*. 2010;66(7):o1835–6.

28. Girgis AS, Hosni HM, Ahmed-Farag IS. A convenient regioselective synthesis of 6-amino-2-oxo-3, 5-pyridinedicarbonitriles. *Zeitschrift für Naturforschung B*. 2003;58(7):678-85.
29. Jansen RJ, de Gelder R, Rowan AE, Scheeren HW, Nolte RJM. Molecular Clips Based on Propanediurea. Exceptionally High Binding Affinities for Resorcinol Guests. *The Journal of Organic Chemistry*. 2001;66(8):2643-53.
30. Flemig I. *Molecular Orbitals and Organic Chemical Reactions*. John Wiley & Sons Ltd: London, UK; 2009.
31. Vernekar SKV, Hallaq HY, Clarkson G, Thompson AJ, Silvestri L, Lummis SCR, et al. Toward Biophysical Probes for the 5-HT₃ Receptor: Structure–Activity Relationship Study of Granisetron Derivatives. *Journal of Medicinal Chemistry*. 2010;53(5):2324-8.
32. Abu Thaher B, Koch P, Schollmeyer D, Laufer S. 4-(4-Fluorophenyl)-1-phenyl-3-(pyridin-4-yl)-1H-pyrazol-5-amine. *Acta Crystallographica Section E*. 2012;68(3):o632.
33. McMurry JE, Hoeger CA, Peterson VE, Ballantine DS. *Fundamentals of General, Organic, and Biological Chemistry: Pearson New International Edition*: Pearson; 2013:976.
34. Ai X-K, Bi C-F, Fan Y-H, Zhang X, He X-T. 2-[(2-Aminophenylimino)(phenyl)methyl]-4-chlorophenol. *Acta Crystallographica Section E*. 2006;62(8):o3475-o6.
35. Zhi-Ping Li, You-Ji Li, Ai-Hua Shi, Yu-Zhu Ouyang. 3-Hydroxy-N'-[(E)-(2-hydroxy-1-naphthyl)methylene]benzohydrazide - methanol (1:1). *Zeitschrift für Kristallographie - New Crystal Structures*. 2008;223(3):293

36. Lionetti D, Medvecz AJ, Ugrinova V, Quiroz-Guzman M, Noll BC, Brown SN. Redox-Active Tripodal Aminetris(aryloxide) Complexes of Titanium(IV). *Inorganic Chemistry*. 2010;49(10):4687-97.
37. Yu L, Hao W, Heng-Yi Z, Bang-Tun Z, Li-Hua W. Novel linear molecular aggregation tethered by hydrogen-Bonded interaction within the crystalline calix[4]arene derivatives. *Journal of Supramolecular Chemistry*. 2002;2(6):515 - 9.
38. Smith CB, Buntine MA, Lincoln SF, Taylor MR, Wainwright KP. Structure of the Molecular Receptor 1,4,7,10-Tetrakis[(S)-2-hydroxy-2-phenylethyl]-1,4,7,10-tetraazacyclododecane: A Combined X-Ray Crystallographic and Theoretical Study Producing an Assessment of the Crystal Packing Energy. *Australian Journal of Chemistry*. 2006;59(2):123-8.
39. Shastin AV, Godovikova TI, Pivina TS, Golovina NI, Shilov GV, Strelenko YA, et al. The structure of 2,4,6-tris[di(tert-butoxycarbonyl)methylidene]-hexahydro-1,3,5-triazine. *Russian Chemical Bulletin (Translation of Izvestiya Akademii Nauk, Seriya Khimicheskaya)*. 2006;55(6):1060-5.
40. Bruno G, Cafeo G, Kohnke FH, Nicolò F. Tuning the anion binding properties of calixpyrroles by means of p-nitrophenyl substituents at their meso-positions. *Tetrahedron*. 2007;63(40):10003-10.
41. Cafeo G, Kohnke FH, La Torre GL, White AJP, Williams DJ. From Large Furan-Based Calixarenes to Calixpyrroles and Calix[n]furan[m]pyrroles: Syntheses and Structures. *Angewandte Chemie International Edition*. 2000;39(8):1496-8.

42. Stibora I, Růžičková M, Krátký R, Vindyša M, Havlíček J, Pinkhassik E, et al. New Calix[4]arene-Based Amides - Their Synthesis, Conformation, Complexation. Collection of Czechoslovak Chemical Communications. 2001;66(4):641.
43. Kliachyna MA, Yesypenko OA, Pirozhenko VV, Shishkina SV, Shishkin OV, Boyko VI, et al. Synthesis, optical resolution and absolute configuration of inherently chiral calixarene carboxylic acids. Tetrahedron. 2009;65(34):7085-91.
44. Arnaud-Neu F, Barbosa S, Fanni S, Schwing-Weill M-J, McKee V, McKervery MA. Alkali and Alkaline Earth Ion Complexation and X-ray Crystal Structure of p-tert-Butylcalix[4]arene Tetraethylamide. Industrial & Engineering Chemistry Research. 2000;39(10):3489-92.
45. Guelzim A, Khrifi S, Baert F, Saadioui M, Asfari Z, Vicens J. 1,3-Di(ethoxy-ethoxy-methoxy)calix[4]arene. Acta Crystallographica Section C. 1997;53(12):1958-60.
46. Schramm H, Saak W, Hoenke C, Christoffers J. Synthesis of Triazolyl-Substituted 3-Aminopiperidines by Huisgen-1,3-Dipolar Cycloaddition – New Scaffolds for Combinatorial Chemistry. European Journal of Organic Chemistry. 2010;2010(9):1745-53.
47. Aversa MC, Barattucci A, Bonaccorsi P, Faggi C, Papalia T. Thiacyclophane Cages and Related Bi- and Tripodal Molecules via Transient Polysulfenic Acids. The Journal of Organic Chemistry. 2007;72(12):4486-96.
48. Krill J, Shevchenko IV, Fischer A, Jones PG, Schmutzler R. 1,5-dimethyl-2,3,3,4-tetrachloro-1,5,2,4-diazadiphosphorinan-6-one and some derivatives. Part II. Heteroatom Chemistry. 1997;8(2):165-75.

49. Khan FA, Sudheer C. A Chiral Pool Approach to the Synthesis of Optically Active Tetrahalo Norbornyl Building Blocks. *Organic Letters*. 2008;10(14):3029-32.
50. Yamamoto Y, Takagishi H, Itoh K. Ruthenium(II)-Catalyzed [2 + 2 + 2] Cycloaddition of 1,6-Diynes with Tricarbonyl Compounds. *Journal of the American Chemical Society*. 2002;124(24):6844-5.
51. Van Rossom W, Caers J, Robeyns K, Van Meervelt L, Maes W, Dehaen W. (Thio)ureido Anion Receptors Based on a 1,3-Alternate Oxacalix[2]arene[2]pyrimidine Scaffold. *The Journal of Organic Chemistry*. 2012;77(6):2791-7.
52. Kanlayavattanakul M, Ruangrunsi N, Watanabe T, Kawahata M, Therrien B, Yamaguchi K, et al. ent-Halimane Diterpenes and a Guaiane Sesquiterpene from *Cladogynos orientalis*. *Journal of Natural Products*. 2005;68(1):7-10.
53. Schütz J, Windisch P, Kristeva E, Wurst K, Ongania K-H, Horvath UEI, et al. Mechanistic Diversity of the van Leusen Reaction Applied to 6-Ketomorphinans and Synthetic Potential of the Resulting Acrylonitrile Substructures. *The Journal of Organic Chemistry*. 2005;70(13):5323-6.
54. Yu H, Li J, Kou Z, Du X, Wei Y, Fun H-K, et al. Photoinduced Tandem Reactions of Isoquinoline-1,3,4-trione with Alkynes To Build Aza-polycycles. *The Journal of Organic Chemistry*. 2010;75(9):2989-3001.
55. Pirrung MC. Ethylene biosynthesis. 8. Structural and theoretical studies. *The Journal of Organic Chemistry*. 1987;52(19):4179-84.

56. Infantes L, Fabian L, Motherwell WDS. Organic crystal hydrates: what are the important factors for formation. *CrystEngComm*. 2007;9(1):65-71.
57. Choi S, Silverman RB. Inactivation and Inhibition of γ -Aminobutyric Acid Aminotransferase by Conformationally Restricted Vigabatrin Analogues. *Journal of Medicinal Chemistry*. 2002;45(20):4531-9.
58. Görbitz CH. Solvent Site Preferences in the Crystal Structures of L-Leucyl-L-leucine Alcohol (1: 1) Complexes. *Acta Chemica Scandinavica*. 1998;52:1343-9.
59. Görbitz CH, Torgersen E. Symmetry, pseudosymmetry and packing disorder in the alcohol solvates of L-leucyl-L-valine. *Acta Crystallographica Section B*. 1999;55(1):104-13.
60. Dreger A, Münster N, Nieto-Ortega B, Ramírez FJ, Gjika M, Schmidta A. Pyrazolium-sulfonates. Mesomeric betaines possessing iminium-sulfonate partial structures. *Archive for Organic Chemistry*. 2012;3:20-37.
61. Takahashi A, Nakamura H, Ikeda D, Naganawa H, Kameyama T, Kurasawa S, et al. Thrazarine, a new antitumor antibiotic. II. Physico-chemical properties and structure determination. *The Journal of antibiotics*. 1988;41(11):1568-74.
62. Fursova EY, Ovcharenko VI, Romanenko GV, Tretyakov EV. A new method for the reduction of nitronyl nitroxides. *Tetrahedron Letters*. 2003;44(34):6397-9.
63. Todorov P, Calmes M, Shivachev BL, Nikolova RP. (R)-Methyl {[2-carboxybicyclo[2.2.2]octan-1-yl]ammonio}methyl}phosphonate dichloromethane 0.25-solvate. *Acta Crystallographica Section E*. 2011;67(8):o2152-o3.

64. Szablewski M, Thomas PR, Thornton A, Bloor D, Cross GH, Cole JM, et al. Highly Dipolar, Optically Nonlinear Adducts of Tetracyano-p-quinodimethane: Synthesis, Physical Characterization, and Theoretical Aspects. *Journal of the American Chemical Society*. 1997;119(13):3144-54.
65. Maas G. Methyl 3,5-di-tert-butyl-4-diisopropylamino-1H⁺-pyridazinium-6-phosphonate dichloromethane solvate, C₁₉H₃₆N₃O₃P·0.85 CH₂Cl₂: a betaine having a nonplanar pyridazinium ring. *Acta Crystallographica Section C*. 1985;41(7):1130-3.
66. Schaumann E, Behr H, Adiwidjaja G. Zur Chemie der α-Thiocarbamoylcarbodiimide. *Liebigs Annalen der Chemie. European Journal of Organic Chemistry*. 1979;1979(9):1322-36.
67. Betzl W, Hettstedt C, Karaghiosoff K. New anellated 4H-1,4,2-diazaphospholes. *New Journal of Chemistry*. 2013;37(2):481-7.
68. Siemeling U, Memczak H, Bruhn C, Vogel F, Trager F, Baio JE, et al. Zwitterionic dithiocarboxylates derived from N-heterocyclic carbenes: coordination to gold surfaces. *Dalton Transactions*. 2012;41(10):2986-94.
69. Shamala N, Row TNG, Venkatesan K. Crystal and molecular structure of allo-4-hydroxy-L-proline dihydrate. *Acta Crystallographica Section B*. 1976;32(12):3267-70.
70. Rømming C, Uggerud E. The crystal and molecular structure of 8-hydroxy-1-methylquinolinium chloride hydrate and 1-methylquinolinium-8-olate dihydrate. *Acta Chemica Scandinavica Series B Organic Chemistry and Biochemistry*. 1983;37(9):791-5.
71. Bailey NA, Newton CG. 2-Methylaminopyridinium Dinitromethylide Monohydrate, C₇H₈N₄O₄·H₂O. *Crystal Structure Communications*. 1980;9(1):49-56.

72. Abraham MH, Platts JA. Hydrogen bond structural group constants. *The Journal of Organic Chemistry*. 2001;66(10):3484-91.
73. Hunter CA. Quantifying intermolecular interactions: guidelines for the molecular recognition toolbox. *Angewandte Chemie International Edition*. 2004;43(40):5310-24.
74. Balducci D, Lazzari I, Monari M, Piccinelli F, Porzi G. (S)- α -methyl, α -amino acids: a new stereocontrolled synthesis. *Amino Acids*. 2010;38(3):829-37.
75. Mangion D, Arnold DR, Cameron TS, Robertson KN. The electron transfer photochemistry of allenes with cyanoarenes. Photochemical nucleophile-olefin combination, aromatic substitution (photo-NOCAS) and related reactions. *Journal of the Chemical Society, Perkin Transactions 2*. 2001(1):48-60.
76. Etter MC. Encoding and decoding hydrogen-bond patterns of organic compounds. *Accounts of Chemical Research*. 1990;23(4):120-6.
77. Galek PT, Fábíán L, Motherwell WS, Allen FH, Feeder N. Knowledge-based model of hydrogen-bonding propensity in organic crystals. *Acta Crystallographica Section B: Structural Science*. 2007;63(5):768-82.

Chapter 7: Experimental validation of the models

7.1 Overview

The predictive models for the five solvents were fitted based on the information available in the CSD. The performance of the models was statistically evaluated and the data misclassified by the model was analysed in previous chapters. At this point, it was important to check the applicability of these models in real-life experiments, representing their use in pharmaceutical development. In this chapter, 10 pharmaceutically active candidates will be selected. They will be experimentally tested for hydrate and solvate formation with each of the five solvents, in order to validate the predictive models obtained. Some of these drugs were part of the training data of solvents. Strictly speaking, this could cause preferential prediction for these candidates. Nevertheless, due to the large number of molecules in each dataset, the effect of these entries on the models is minute, where they constitute less than 1.7 % of the training data in the worst case.

7.2 Selection of drug candidates and their profiles

The 10 drug candidates that were used for the models validation were chosen to be pharmaceutically active and show variability in their molecular structures. The variability among these candidates was in terms of the molecular properties that were considered in the two-variable models. These are molecular size, different level of branching as well as different hydrogen bonding capability. Validating the models based on such different candidates would cover a wide range of pharmaceutical applications. The 10 candidates, their structure and selection strategy are summarized in Table 7-1.

Table 7-1. The molecular structure, some selected molecular descriptors, application and selection strategy of the 10 drug candidates

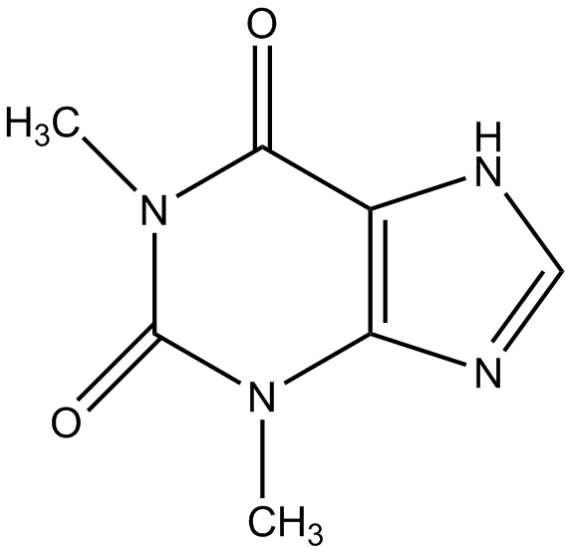
Drug Candidate	Chemical structure								
Theophylline anhydrous									
-	Descriptor values								
-	AVS_H2	TRS	nHDon	SM3_H2	Hy	H.050	piID	nH	
-	3.522	11	1	3.853	0.022	1	6.794	8	
-	Pharmaceutical use								
-	<p>A bronchodilator and an anti-inflammatory that is used in the treatment of reversible respiratory airways obstruction.^{1, 2} Theophylline is one of the most widely used bronchodilators.¹ It has been reported^{3, 4} that low doses of theophylline can reverse corticosteroid resistance in chronic asthma and obstructive pulmonary disease. It also exhibits immunomodulatory effect.³⁻⁵</p>								
-	Selection Philosophy								
-	<p>Theophylline is a small-sized structure as compared to the datasets provided, when the molecular weight is considered (see Figure 4-11 of chemical space in chapter 4) is considered. The structure possesses some short branches. In terms of hydrogen bonding, the structure possesses one hydrogen bond donor and 5 Hydrogen bond acceptors. This candidate was chosen due to the mixed properties it possesses where it falls in the middle range of most descriptors that are important for solvate formation.</p>								

Table 7-1. Continued

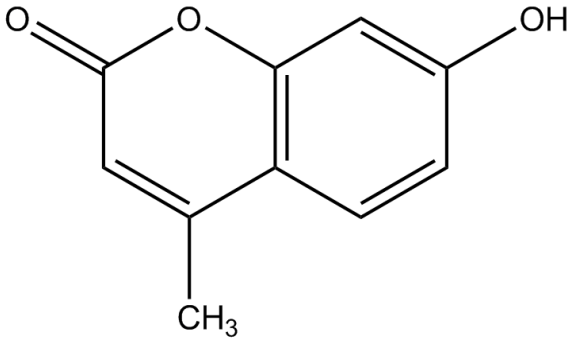
Drug Candidate	Chemical structure									
Hymecromone	 <chem>CC1=C(C(=O)O1)c2ccccc2O</chem>									
-	Descriptor values									
-	AVS_H2	TRS	nHDon	SM3_H2	Hy	H.050	piID	nH		
-	3.443	12	1	3.768	-0.202	1	7.351	8		
-	Pharmaceutical use									
-	A food supplement that is used in bile therapy. ⁶ More recently, it is being recognized as a Hyaluronan (HA, Hyaluronanic acid) inhibitor therefore capable of preventing chronic inflammation, autoimmunity and tumours. ⁷ For example, the inhibitory action of 4-MU on HA can be beneficial in different malignant diseases such as prostate cancer. ⁸									
-	Selection Philosophy									
-	Similar to theophylline, this molecule is small in terms of size, possesses two fused rings and shows minimal branching. It also contains one hydrogen bond donor and 3 acceptors. It is expected to have similar behaviour to theophylline.									

Table 7-1. Continued

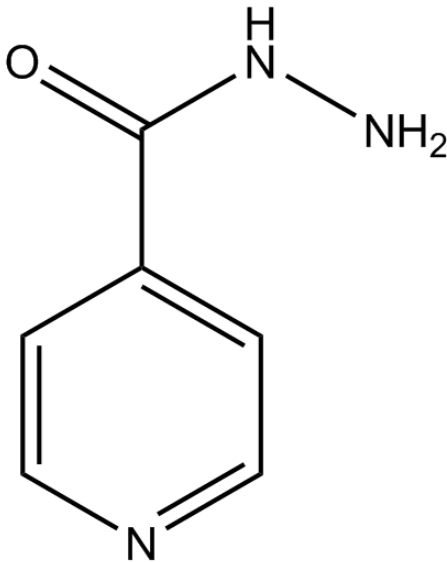
Drug Candidate	Chemical structure																
Isoniazid																	
-	Descriptor values																
-	<table><tr><td>AVS_H2</td><td>TRS</td><td>nHDon</td><td>SM3_H2</td><td>Hy</td><td>H.050</td><td>piID</td><td>nH</td></tr><tr><td>2.97</td><td>6</td><td>3</td><td>3.232</td><td>1.786</td><td>3</td><td>5.833</td><td>7</td></tr></table>	AVS_H2	TRS	nHDon	SM3_H2	Hy	H.050	piID	nH	2.97	6	3	3.232	1.786	3	5.833	7
AVS_H2	TRS	nHDon	SM3_H2	Hy	H.050	piID	nH										
2.97	6	3	3.232	1.786	3	5.833	7										
-	Pharmaceutical use																
-	Isoniazid is a first-line anti-tubercular drug also used in tuberculosis preventive therapy for HIV positive patients. ^{9, 10} Isoniazid is mostly used in combination with other anti-tubercular drugs (ethambutol, rifampicin, and pyrazinamide) to treat multi-drug resistant tuberculosis. ^{9, 11}																
-	Selection Philosophy																
-	Isoniazid possesses an AVS_H2 value which is well below the average value in any of the datasets. On the other hand it possesses 3 hydrogen bond donors in addition to 4 hydrogen bond acceptors, all of which are accessible for interaction. This hydrogen bonding ability is also well above the average, compared to other entries in any dataset (see Figure 4-11 from chapter 4 for more details). Such properties make isoniazid an excellent candidate that can compare the importance of the size, branching and complexity of the structure against the importance of hydrogen bonding.																

Table 7-1. Continued

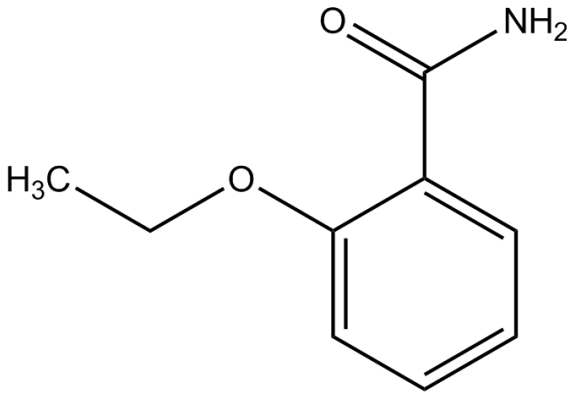
Drug Candidate	Chemical structure									
Ethenzamide										
-	Descriptor values									
-	AVS_H2	TRS	nHDon	SM3_H2	Hy	H.050	piID	nH		
-	3.105	6	2	3.477	0.59	2	6.278	11		
-	Pharmaceutical use									
-	<p>Ethenzamide is an anti-inflammatory drug with analgesic, antipyretic, and sedative effects.¹² It also has anti-convulsive and muscle relaxing effects.¹³ Ethenzamide is used as an active ingredient in pain reliving ointments.¹⁴ This drug has been reported to have ulcerogenic and hypothermic effects.¹²</p>									
-	Selection Philosophy									
-	<p>Ethenzamide is structurally related to isoniazid. In terms of size and branching, it possesses an AVS_H2 value that is slightly higher than isoniazid due to the extra branching from the ring. In terms of hydrogen bonding, it shows an inferior hydrogen bond donating/accepting ability. This candidate was chosen to observe if these changes in molecular structure can alter the solvation behaviour of the compound.</p>									

Table 7-1. Continued

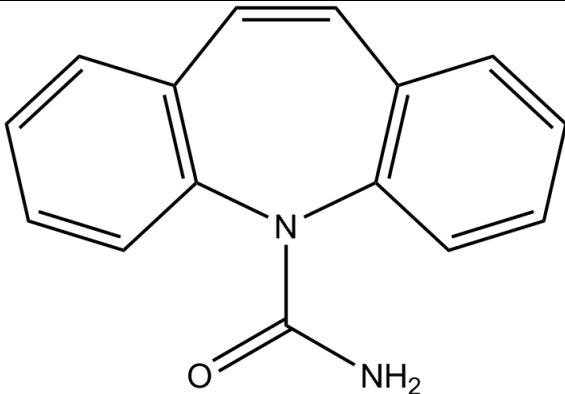
Drug Candidate	Chemical structure																
Carbamazepine																	
-	Descriptor values																
-	<table><tr><th>AVS_H2</th><th>TRS</th><th>nHDon</th><th>SM3_H2</th><th>Hy</th><th>H.050</th><th>piID</th><th>nH</th></tr><tr><td>3.783</td><td>19</td><td>2</td><td>4.168</td><td>0.32</td><td>2</td><td>9.776</td><td>12</td></tr></table>	AVS_H2	TRS	nHDon	SM3_H2	Hy	H.050	piID	nH	3.783	19	2	4.168	0.32	2	9.776	12
AVS_H2	TRS	nHDon	SM3_H2	Hy	H.050	piID	nH										
3.783	19	2	4.168	0.32	2	9.776	12										
-	Pharmaceutical use																
-	Carbamazepine is an anticonvulsant used to treat epilepsy ^{15, 16} and trigeminal neuralgia. ¹⁷ It has been shown to have prophylactic and improving effect in manic-depressive illness. ¹⁸⁻²⁰ It has been reported to reduce symptoms of diabetic neuropathy ²¹ and post-traumatic stress disorder. ²²																
-	Selection Philosophy																
-	Carbamazepine's AVS_H2 value is rather high compared to other entries. In terms of hydrogen bonding ability, one primary amide is present. It is also important to notice the rigidity of this structure, despite it is large size. Carbamazepine was chosen after isoniazid and ethezamide, to see the change in behaviour as the candidate gains more size, retains rigidity and gets a lower hydrogen bonding ability.																

Table 7-1. Continued

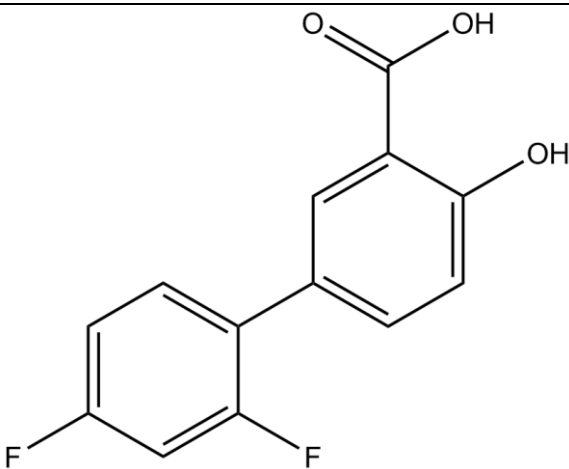
Drug Candidate	Chemical structure									
Diflunisal										
-	Descriptor values									
-	AVS_H2	TRS	nHDon	SM3_H2	Hy	H.050	piID	nH		
-	3.545	12	2	4.083	0.429	2	8.499	8		
-	Pharmaceutical use									
-	Diflunisal is a nonsteroidal anti-inflammatory analgesic. ²³ It has been shown to be efficient in treatment of postoperative pain. ^{24,25,26}									
-	Selection Philosophy									
-	In terms of its molecular structure, diflunisal is considerably different from the previously described candidates. Firstly, it is a flexible compound compared to the previously chosen candidates, where the two rings are connected by a rotatable bond. Secondly, it possesses two fluorine atoms, which could have significant effect on the electrostatic interactions of the molecule. Finally, this entry still has two hydrogen bond donors that are completely available for interacting with other molecules. This drug is a candidate that represents a mildly flexible structure with hydrogen donating ability.									

Table 7-1. Continued

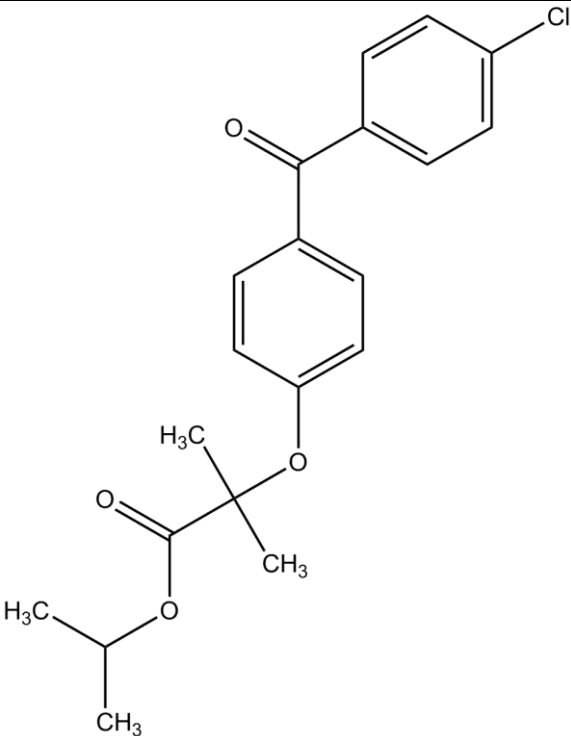
Drug Candidate	Chemical structure																
Fenofibrate																	
-	Descriptor values																
-	<table><tr><td>AVS_H2</td><td>TRS</td><td>nHDon</td><td>SM3_H2</td><td>Hy</td><td>H.050</td><td>piID</td><td>nH</td></tr><tr><td>3.579</td><td>12</td><td>0</td><td>4.381</td><td>-0.79</td><td>0</td><td>8.814</td><td>21</td></tr></table>	AVS_H2	TRS	nHDon	SM3_H2	Hy	H.050	piID	nH	3.579	12	0	4.381	-0.79	0	8.814	21
AVS_H2	TRS	nHDon	SM3_H2	Hy	H.050	piID	nH										
3.579	12	0	4.381	-0.79	0	8.814	21										
-	Pharmaceutical use																
-	Fenofibrate is used to treat high cholesterol levels. ²⁷ It reduces the angiographic progression of coronary-artery disease, ²⁸ #873], #873], #873], #873], #873] 2001 #873;Simpson, 1990 #867] and diabetic retinopathy ²⁹ in diabetes patients.																
-	Selection Philosophy																
-	Fenofibrate structure is largely flexible which gives the freedom to the molecule to assume a variety of conformations. On the other hand, it does not show long branching in any direction. No hydrogen bond donors are available in this structure. This could be a good candidate to see the effect of having a flexible structure with short branches, no hydrogen bond donating ability and a larger size compared to other candidates.																

Table 7-1. Continued

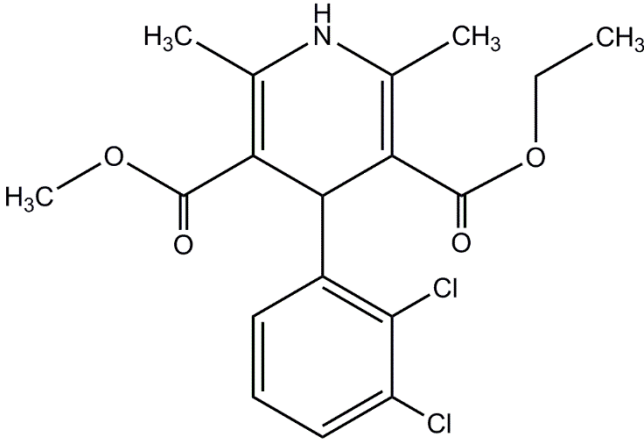
Drug Candidate	Chemical structure																
Felodipine																	
-	Descriptors value																
-	<table><tr><th>AVS_H2</th><th>TRS</th><th>nHDon</th><th>SM3_H2</th><th>Hy</th><th>H.050</th><th>piID</th><th>nH</th></tr><tr><td>3.8</td><td>12</td><td>1</td><td>4.462</td><td>-0.277</td><td>1</td><td>8.66</td><td>19</td></tr></table>	AVS_H2	TRS	nHDon	SM3_H2	Hy	H.050	piID	nH	3.8	12	1	4.462	-0.277	1	8.66	19
AVS_H2	TRS	nHDon	SM3_H2	Hy	H.050	piID	nH										
3.8	12	1	4.462	-0.277	1	8.66	19										
-	Pharmaceutical use																
-	An antihypertensive agent used in the treatment of high blood pressure ³⁰ [Comparison of Antihypertensive Effect and Pharmacokinetics of Conventional and Extended Release Felodipine Tablets in Patients with Arterial Hypertension Drugs]. The effect of Felodipine is increased by combining it with other agents, such as metoprolol. ³¹																
-	Selection Philosophy																
-	Felodipine has a highly branched structure with a high AVS_H2 value. It also possesses one hydrogen bond donor and multiple hydrogen bond acceptors. The features of this candidate are close to fenofibrate, which might cause them to have similar behaviour.																

Table 7-1. Continued

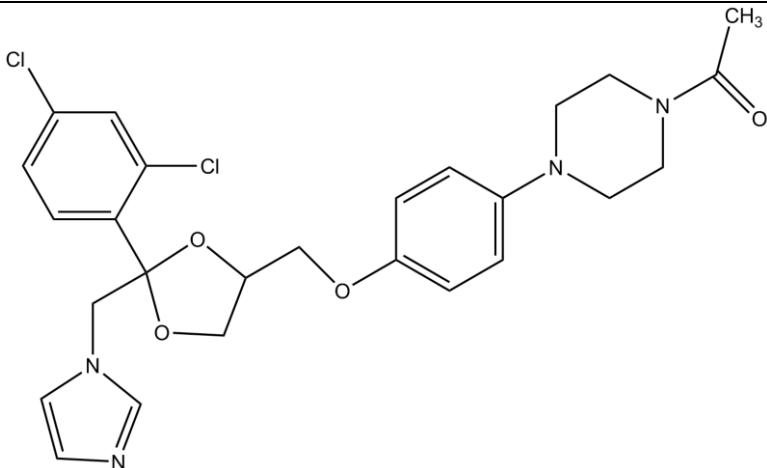
Drug Candidate	Chemical structure																
Ketoconazole																	
-	Descriptor values																
-	<table><tr><td>AVS_H2</td><td>TRS</td><td>nHDon</td><td>SM3_H2</td><td>Hy</td><td>H.050</td><td>piID</td><td>nH</td></tr><tr><td>3.924</td><td>28</td><td>0</td><td>4.852</td><td>-0.717</td><td>0</td><td>10.425</td><td>28</td></tr></table>	AVS_H2	TRS	nHDon	SM3_H2	Hy	H.050	piID	nH	3.924	28	0	4.852	-0.717	0	10.425	28
AVS_H2	TRS	nHDon	SM3_H2	Hy	H.050	piID	nH										
3.924	28	0	4.852	-0.717	0	10.425	28										
-	Pharmaceutical use																
-	An antifungal agent ³² that is used to treat systematic dermal infections. ³³ It can be administered orally or topically (as cream or shampoo). ³⁴ The oral form is not preferred due to its high hepatotoxicity. ³⁴ Higher doses of the drug can cause hypoadrenalism. ³⁵																
-	Selection Philosophy																
-	Ketoconazole represents the case when the size of a molecule is large, branched, possesses multiple hydrogen bond acceptors, but no hydrogen bond donors. Screening this structure would give an idea on the role of hydrogen donating importance in solvate formation.																

Table 7-1. Continued

Drug Candidate	Chemical structure								
Griseofulvin									
-	Descriptor values								
-	AVS_H2	TRS	nHDon	SM3_H2	Hy	H.050	piID	nH	
-	3.987	17	0	4.563	-0.699	0	8.93	17	
-	Pharmaceutical use								
-	<p>An antifungal that is mainly used in the treatment of skin infections such as trichophyton and microsporum.³⁶⁻³⁸ It is administered orally and known to have multiple side effects.³⁹ More recently, it has been reported to be a promising anti-cancer agent.^{40, 41}</p>								
-	Selection Philosophy								
-	<p>Griseofulvin is very close to ketoconazole in terms of the overall size and branching, as indicated by their AVS_H2 values. They also have identical nHDon and H-050 values and they both possess chlorine atoms. On the other hand griseofulvin is much more rigid compared to ketoconazole. This candidate could show the effect of rigidity on solvate and hydrate formation.</p>								

7.3 Sample preparation and characterization

In this project, it is required to mimic the process under which pharmaceutical materials go during processing and manufacturing. These include exposing the candidates to the solvent for a long time to allow it to convert to the solvate form in addition to the stirring, which resembles the mechanical processing of materials in the industrial process. Therefore, slurries of each of the 10 drug candidates, with each of the 5 solvents were prepared, resulting in 50 slurries. The principal method for hydrate and solvate formation is shown in the materials and methods chapter, section 3.2.2.

Characterization of the samples was conducted using TGA. In this work, highly unstable solvates that undergo complete desolvation already below 40 °C were considered as non-solvates. This is because in practice, they are expected to lose the solvent on their own or during the drying process. Heating up to 250 °C was reasoned by the fact that all 10 candidates melt or decompose before reaching 250 °C.⁴²⁻⁵²

A thermogram was obtained for each of the 50 of the slurries prepared according to the method in section 3.2.2. Additionally, starting materials (labelled as 'raw' in figure) were also screened *via* TGA. Therefore; a set of 6 thermograms per drug candidate was obtained. These were then overlaid and analysed. When a new solvate was detected, further analysis took place to confirm the presence of this new form. It is important to mention that the TGA results shown in the sections of this chapter were normalized. This means that the first data point in each plot is considered to be the reference (100 % of weight). The normalization is denoted in the legend of the plots as an asterisk "*". It is also useful to mention that the distance between the points on the curves is 5 °C.

Theophylline anhydrous

The thermograms obtained for the theophylline screening products are shown in Figure 7-1.

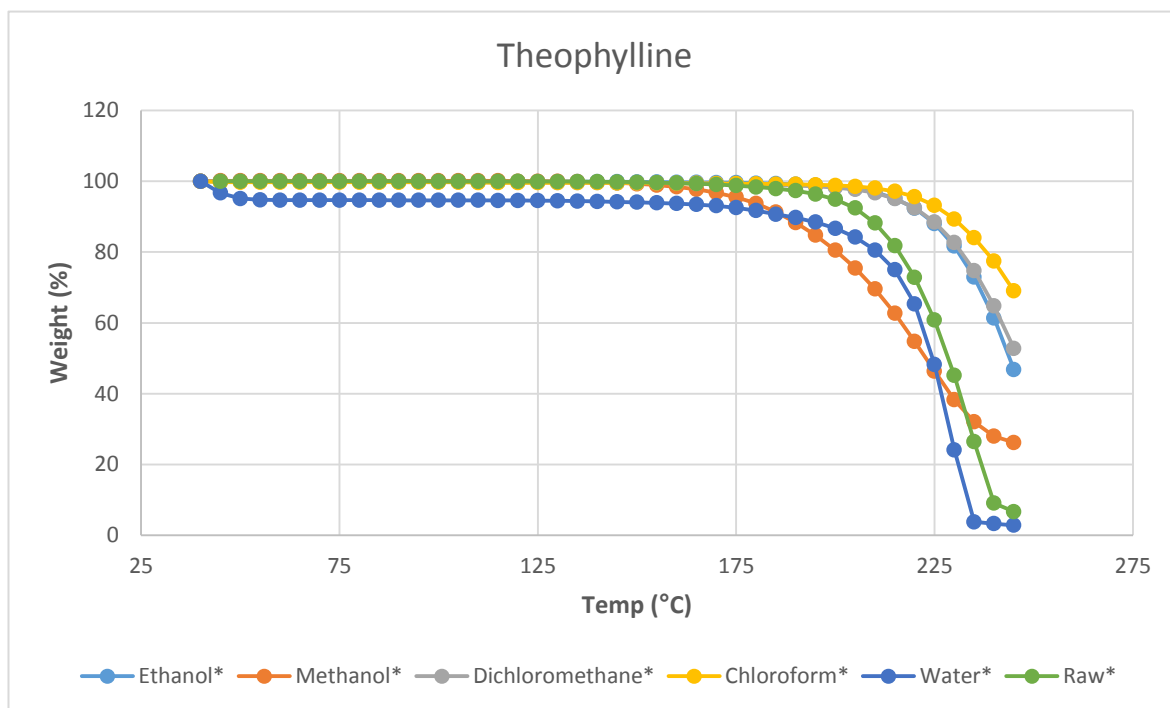


Figure 7-1. The heating profile for theophylline and its screening products.

With the exception of water, no weight loss was observed in any TGA profile prior to decomposition, which is seen above 170 °C. A theophylline hydrate was detected, with a weight loss of ~5.5 % at 60 °C. Nevertheless, the weight loss in this thermogram could be misleading, as the desolvation seems to be starting before 40 °C. Therefore, the TG analysis was repeated from ambient temperature. The result is shown in Figure 7-2.

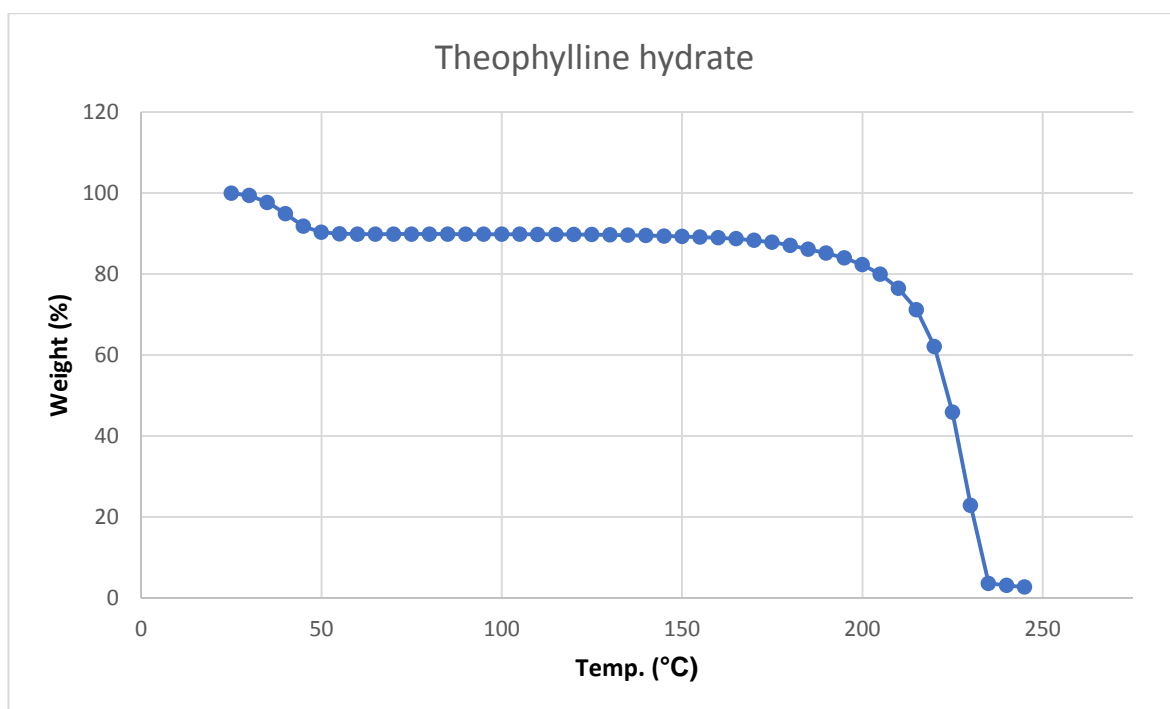


Figure 7-2. The TG thermogram of the theophylline hydrate from room temperature.

The weight loss in Figure 7-2 represents the weight loss starting from room temperature. This means the weight loss observed here is the total loss in the experiment, which was calculated to be 10 % of the total sample weight. Such loss corresponds to a 1:1 stoichiometry of water to theophylline (the theoretical weight loss for a 1:1 hydrate is 9.1 %). Therefore, the hydrate formed here is most likely to be the same hydrate reported by Sutor *et al* In 1958.⁵³ (CSD reference code THEOPH)

Hymecromone

The thermograms obtained for the hymecromone screening products are shown in Figure 7-3.

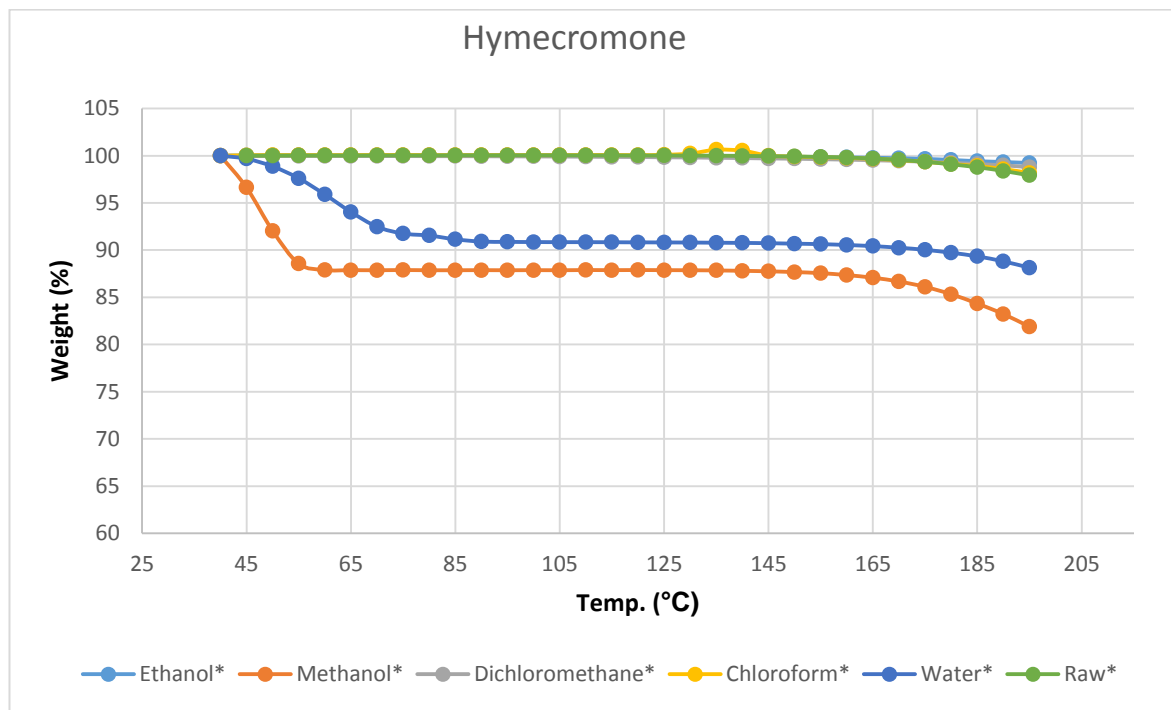


Figure 7-3. TGA profiles of hymecromone and its screening product.

Ethanol, dichloromethane and chloroform were not able to form a solvate with hymecromone. Methanol and water on the other hand have shown a weight drop, attributed to the presence of a solvate in the structure. Since the desolvation that was observed for these two solvate forms starting around 40 °C, there is a chance there has been a weight loss below 40 °C. For this reason, the TGA for these two samples were repeated, with recording the data starting from RT. The results are shown in Figure 7-4.

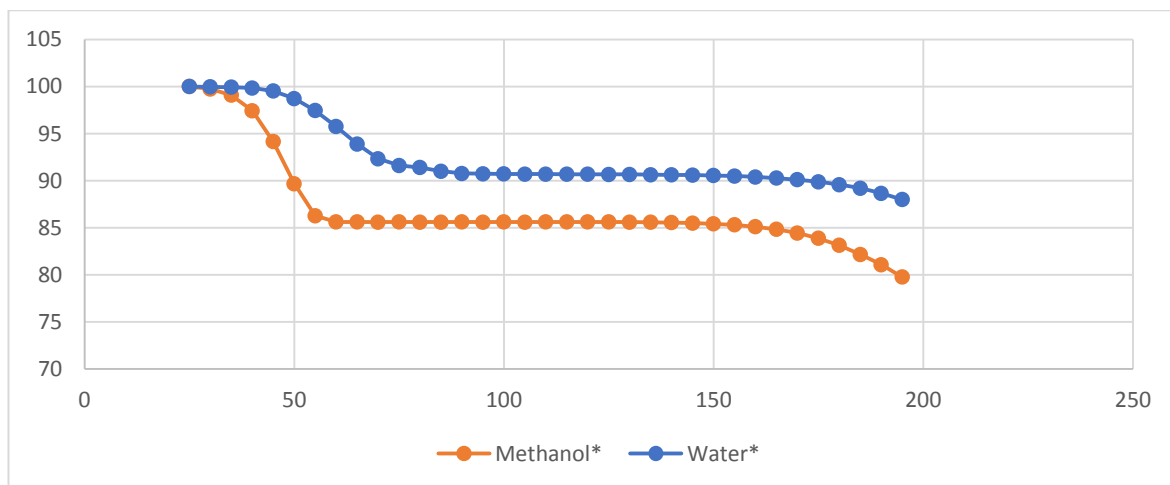


Figure 7-4. TGA profile of hymcromone methanol solvate and hydrate.

The hydrate form shown in Figure 7-4 has reached a plateau of weight loss of 9.2 % at 90 °C, while the methanol solvate reached the weight loss plateau at 60 °C, with a weight loss value of 14.6 % of the original weight. Both of these correspond to 1:1 solvent:drug solvate (the theoretical weight loss is 9.3 % for a 1:1 hydrate and 15.4 % for a 1:1 methanol solvate). A hymcromone hydrate was previously reported in 2002 with the same stoichiometry (1:1) (CSD reference code WIKDAV),⁵⁴ which is likely to be the same hydrate. On the other hand, the methanol solvate was not found to be previously reported. For this reason, the PXRD pattern for the methanol solvate was obtained and compared to the simulated PXRD patterns of other hymcromone forms found in the CSD, as illustrated in Figure 7-5.

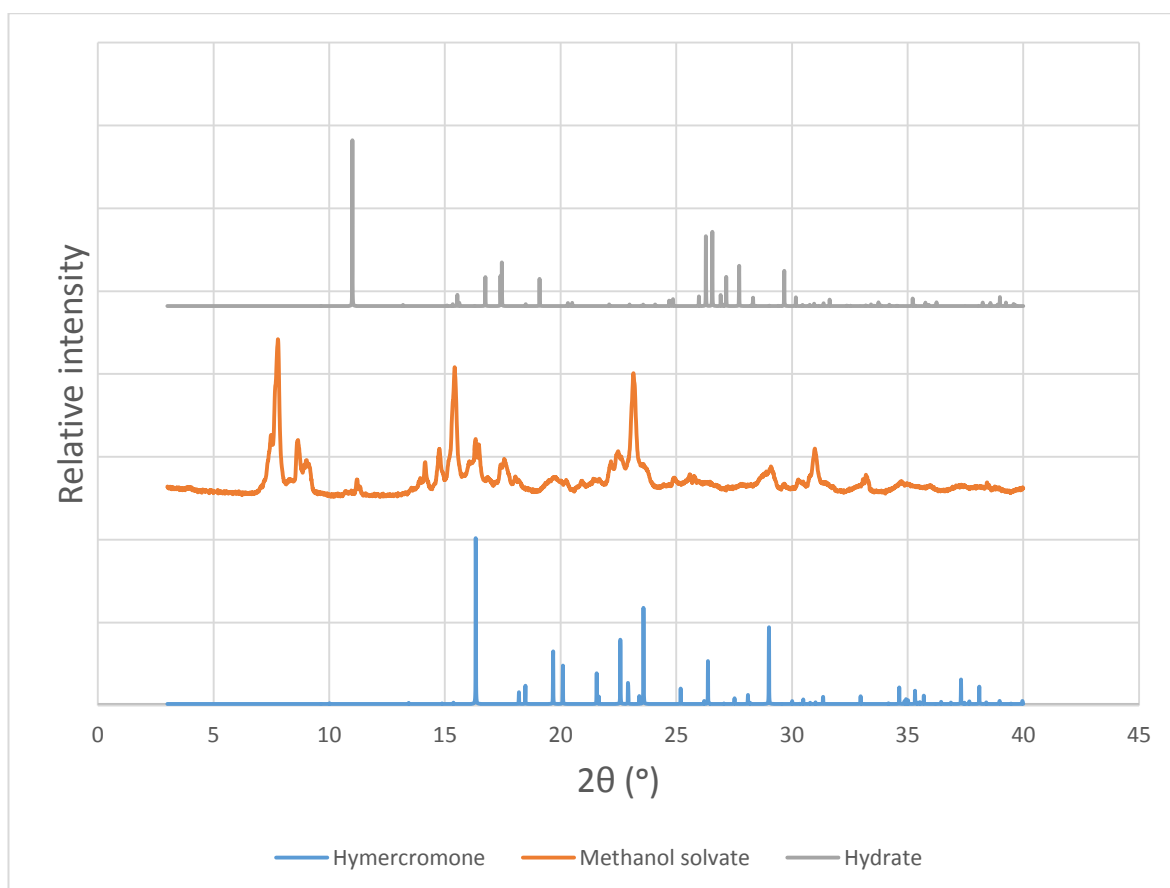


Figure 7-5. PXRD patterns of the solvated forms of hymecromone.

Isoniazid

The thermograms obtained for the isoniazid screening products are shown in Figure 7-6.

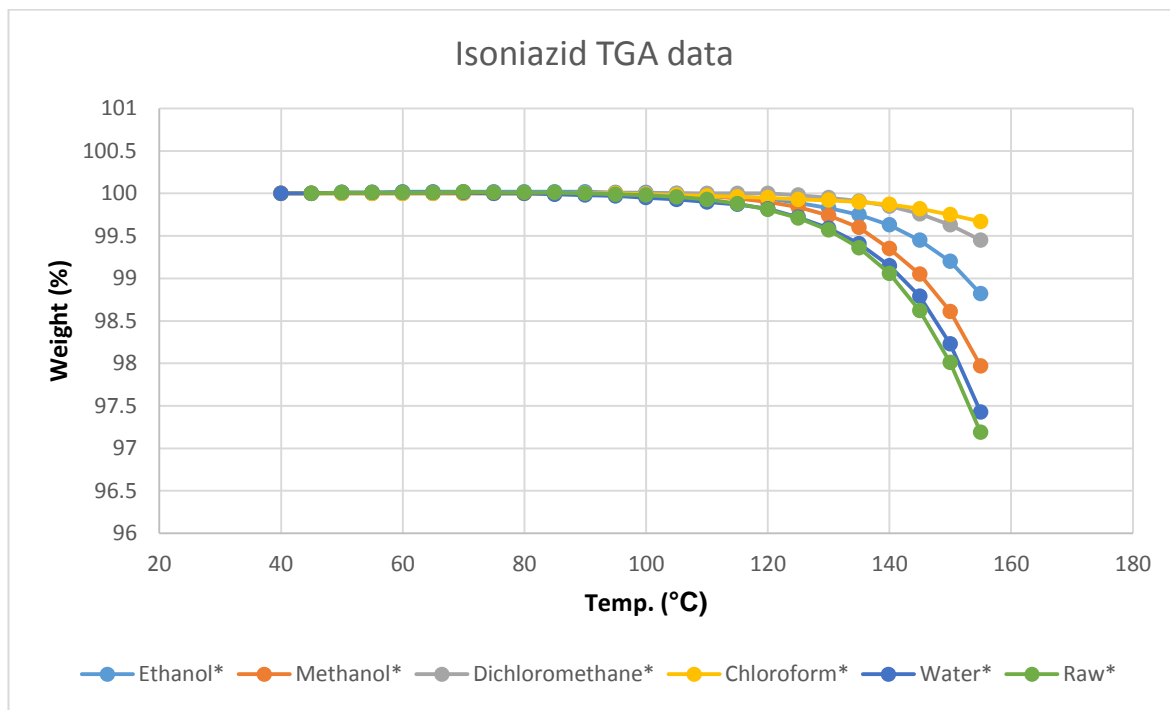


Figure 7-6. TGA profiles of isoniazid and its screening products. Note that the heating was up to 155 °C only; this is due to the low melting point of this drug candidate.

Isoniazid was not able to form a solvate with any of the 5 solvents, where no weight loss was observed in any of the TG plots.

Ethenzamide

The thermograms obtained for the ethenzamide screening products are shown in Figure 7-7.

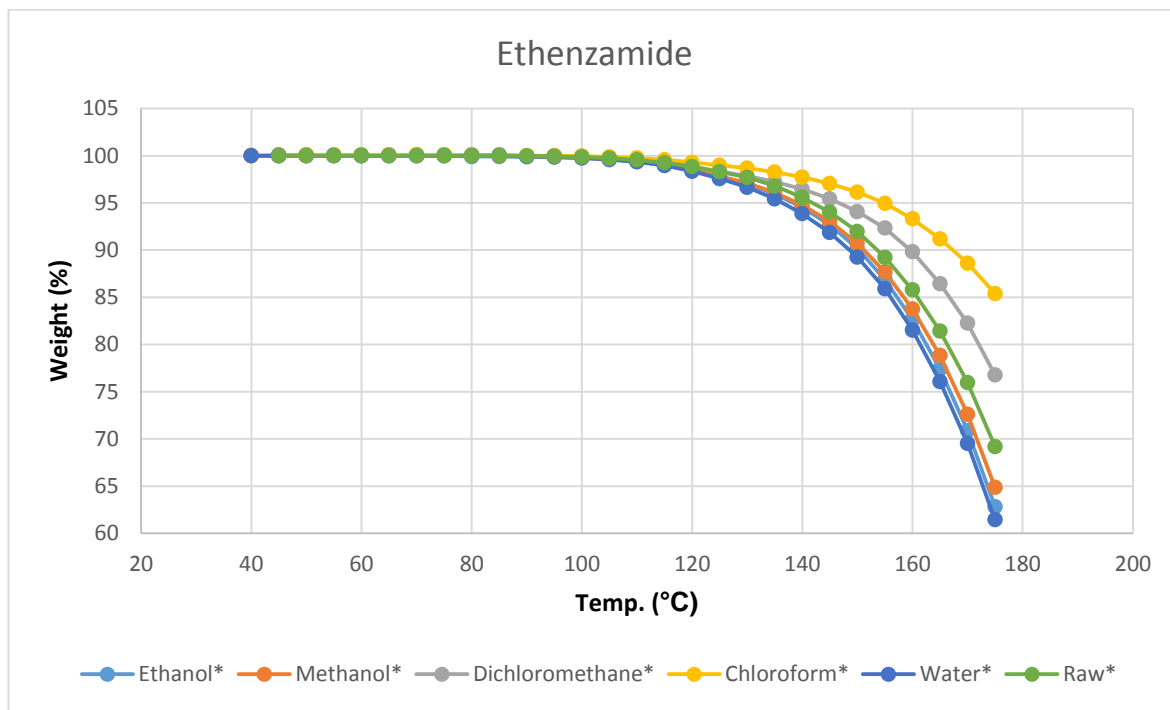


Figure 7-7. TGA profiles of ethenzamide and its screening products. Note that the heating was up to 180 °C only due to the low melting point of this drug candidate.

Similar to isoniazid, ethenzamide did not form a solvate with any of the five solvents. These two drug candidates are structurally related and it is expected for them to show similar behaviour.

Carbamazepine

The thermograms obtained for the carbamazepine screening products are shown in Figure 7-8.

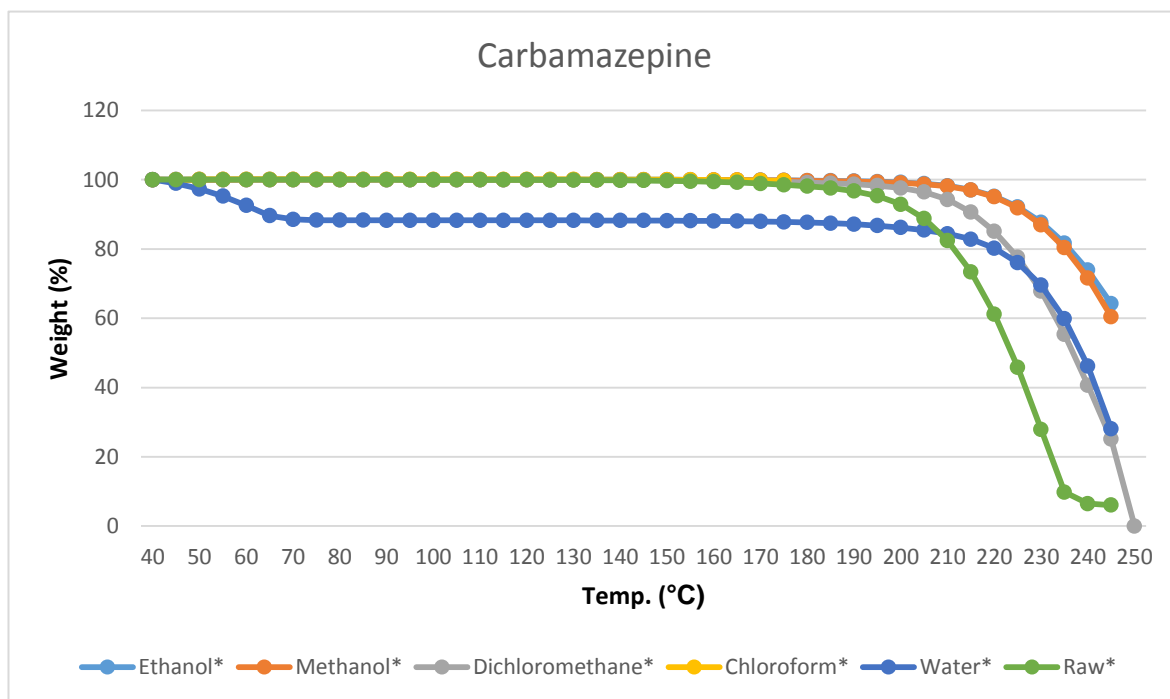


Figure 7-8. TGA profiles of carbamazepine and its screening products.

Carbamazepine has failed to form a solvate with any solvent except for water. A hydrate is already known to form with carbamazepine as has been reported.⁵⁵ As the desolvation of the carbamazepine hydrate takes place at low temperature, the TGA run was repeated starting from RT. The TGA profile obtained is shown in Figure 7-9.

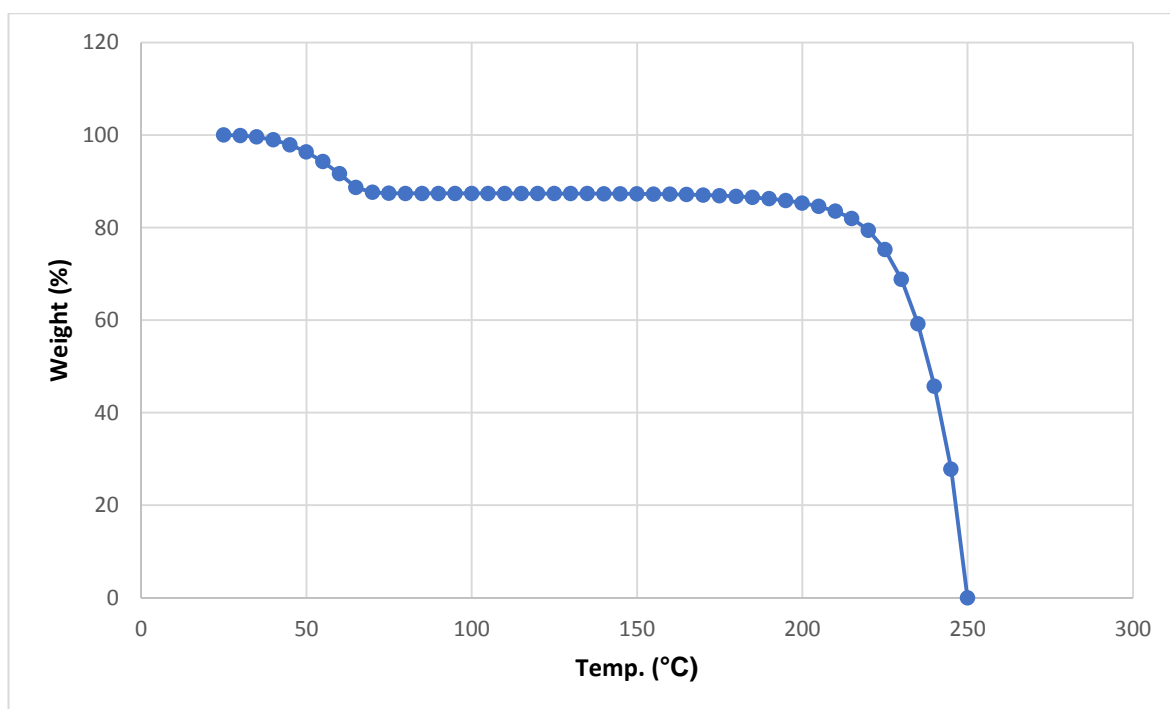


Figure 7-9. TGA profile of carbamazepine hydrate measured from RT.

The weight loss was 12.6 %, this corresponds to a dihydrate, (theoretical weight loss of a dihydrate is 13.3 %) showing that the solvate that was formed here is most likely to be the same as the one known before.

Diflunisal

The thermograms obtained for the diflunisal screening products are shown in Figure 7-10.

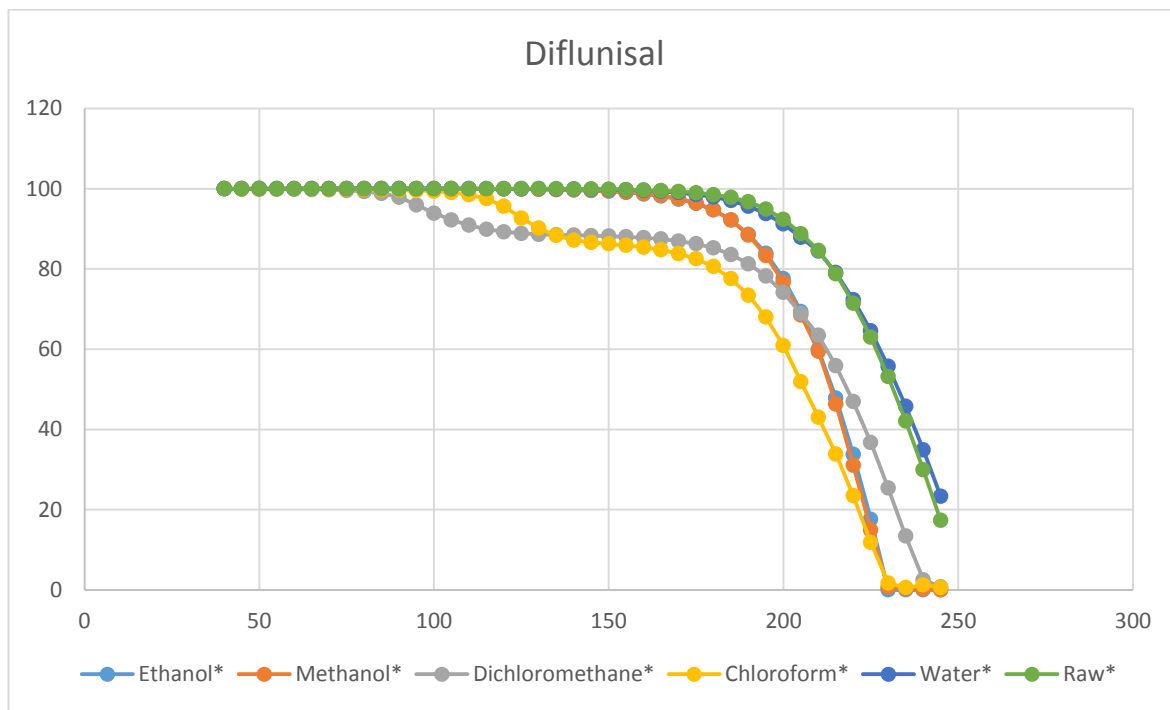


Figure 7-10. TGA profiles of diflunisal and its screening products.

Diflunisal was able to form a solvate with both chlorinated solvents, as can be seen in Figure 7-10. A 1:1 diflunisal chloroform solvate has been previously reported (CSD refcode: RUXRUX 10.1021/cg025589n), while no dichloromethane solvate was found to be reported. The stoichiometry obtained by weight analysis shows that the solvent: drug ratio is 0.34 and 0.38 for dichloromethane and chloroform, respectively. This ratio suggests the chloroform solvate is different from the reported one, as shown by the PXRDs in Figure 7-11.

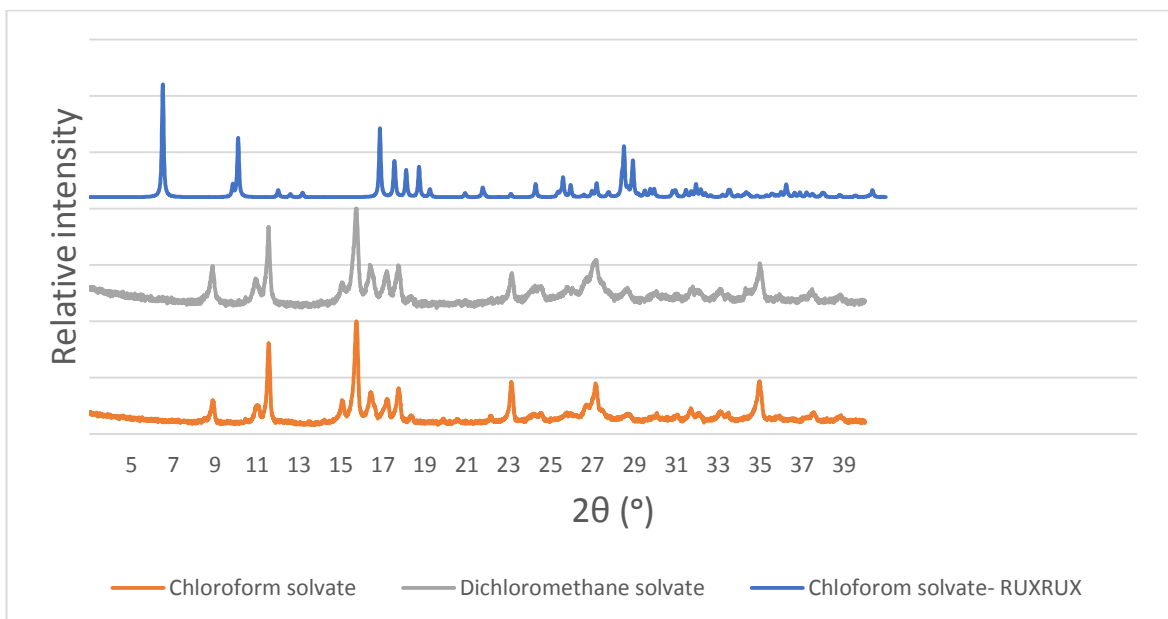


Figure 7-11. PXRD patterns of diflunisal dichloromethane and chloroform solvate forms.

The similarity between the PXRD patterns of the dichloromethane and chloroform solvate and the non-stoichiometric ratio of the solvate forms could be an indicative of the presence of a channel solvate, where the molecules are not in a definite order inside the channels.

Fenofibrate

The thermograms obtained for the fenofibrate screening products are shown in Figure 7-12.

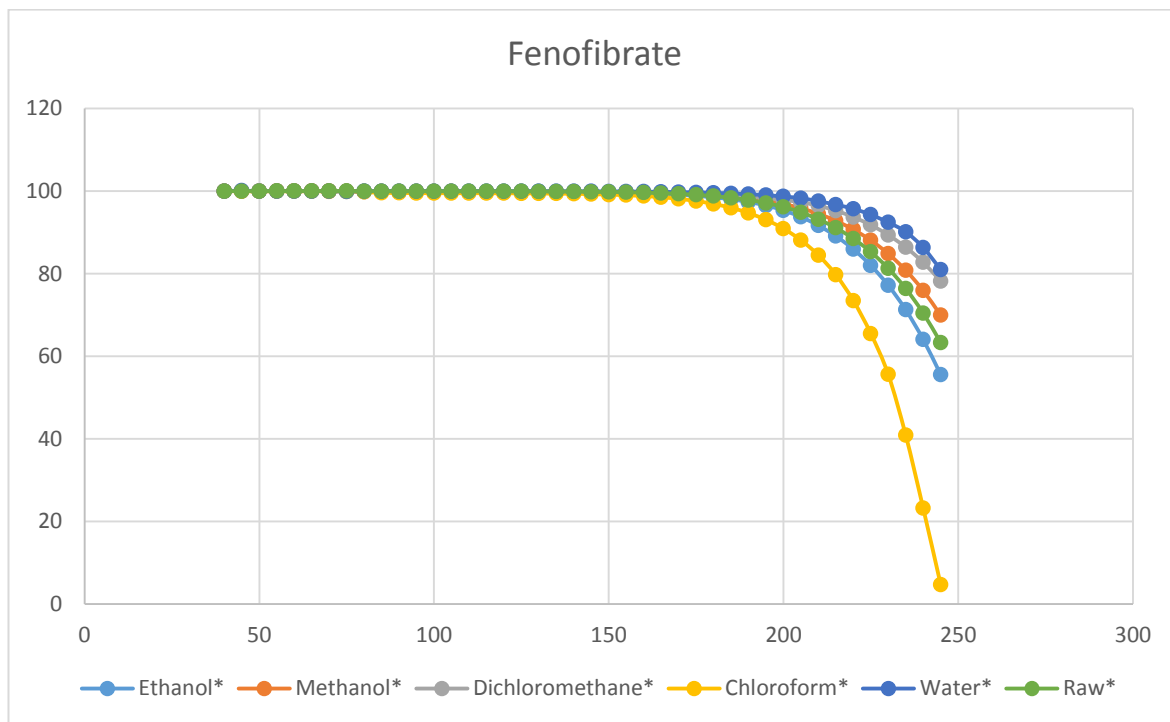


Figure 7-12. TGA profiles of fenofibrate and its screening products.

Despite its large size, fenofibrate has failed to form any solvate with the 5 solvents.

Felodipine

The thermograms obtained for the felodipine screening products are shown in Figure 7-13.

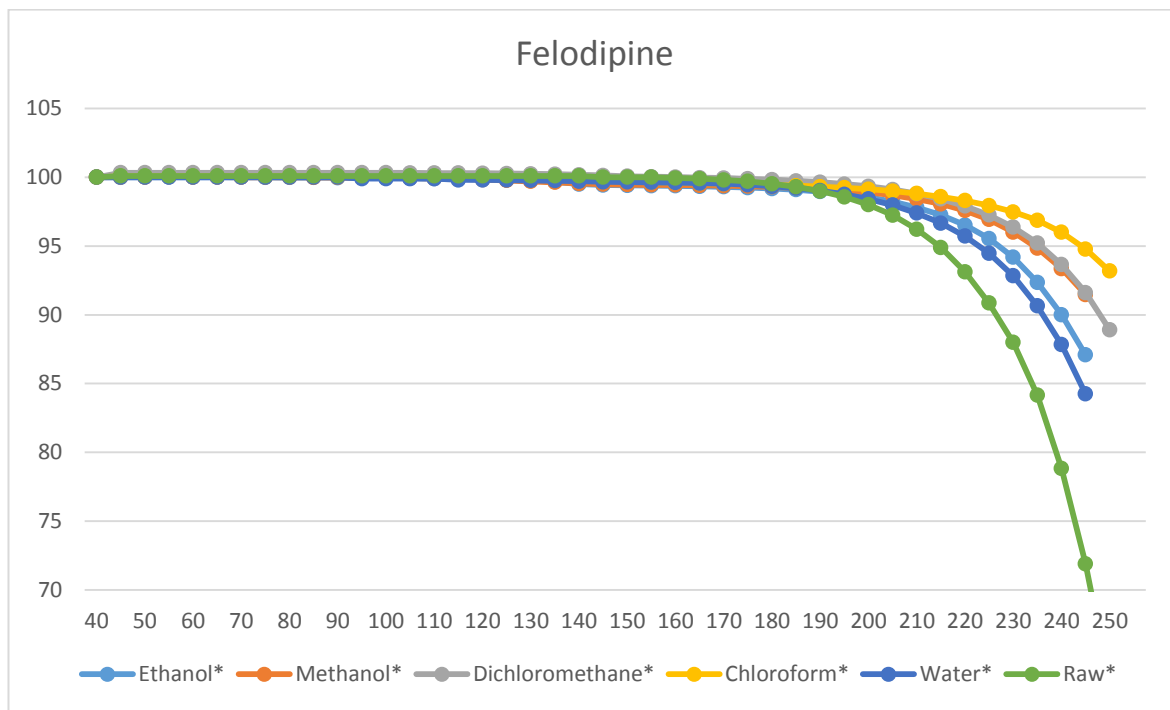


Figure 7-13. TGA profiles of felodipine and its screening products.

Similar to fenofibrate, felodipine is an example of a large-sized structure in comparison to other candidates. Note that felodipine was more likely to form a solvate due to having a hydrogen bond donor in its structure, nevertheless both did not form any solvents with the 5 solvents.

Ketoconazole

The thermograms obtained for the ketoconazole screening products are shown in Figure 7-14.

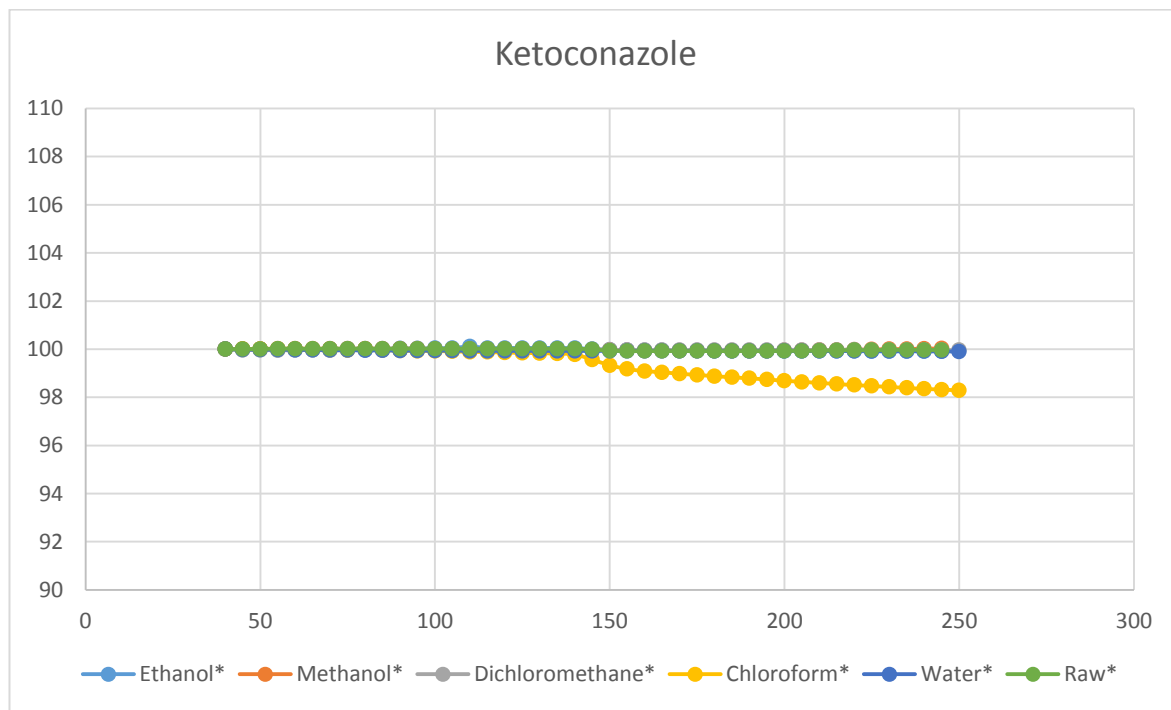


Figure 7-14. TGA profiles of felodipine and its screening products.

Despite its large size, flexibility and multiple hydrogen accepting sites, ketoconazole has failed to form a solvate with any of the 5 solvents. The loss of weight in the chloroform sample has started after melting, when the sample was visually inspected, it was a yellow/brownish liquid in colour, indicating the weight loss seen in the thermogram is due to decomposition.

Griseofulvin

The TGA results for griseofulvin and its screening products are shown in Figure 7-15 .

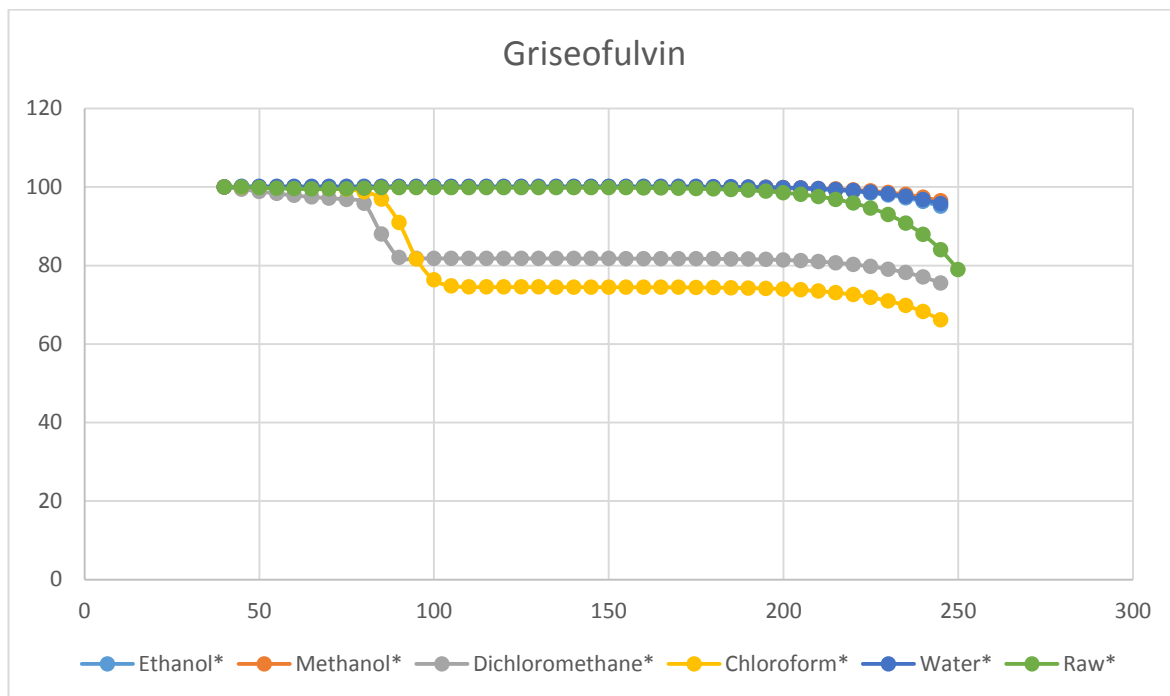


Figure 7-15. TGA profiles of griseofulvin and its screening products.

Griseofulvin was able to form solvates with the chlorinated solvents only. Both a chloroform and a dichloromethane solvates were previously reported for griseofulvin.^{56,57} They show an interesting desolvation pattern, where the solvent loss is observed over two steps, one happening below 80 °C and the other starting around 80 °C. It is noticeable that the solvent loss starts from around room temperature. In order to work out the solvate stoichiometry correctly, another experiment was conducted starting from 25 °C, the result of which is shown in Figure 7-16.

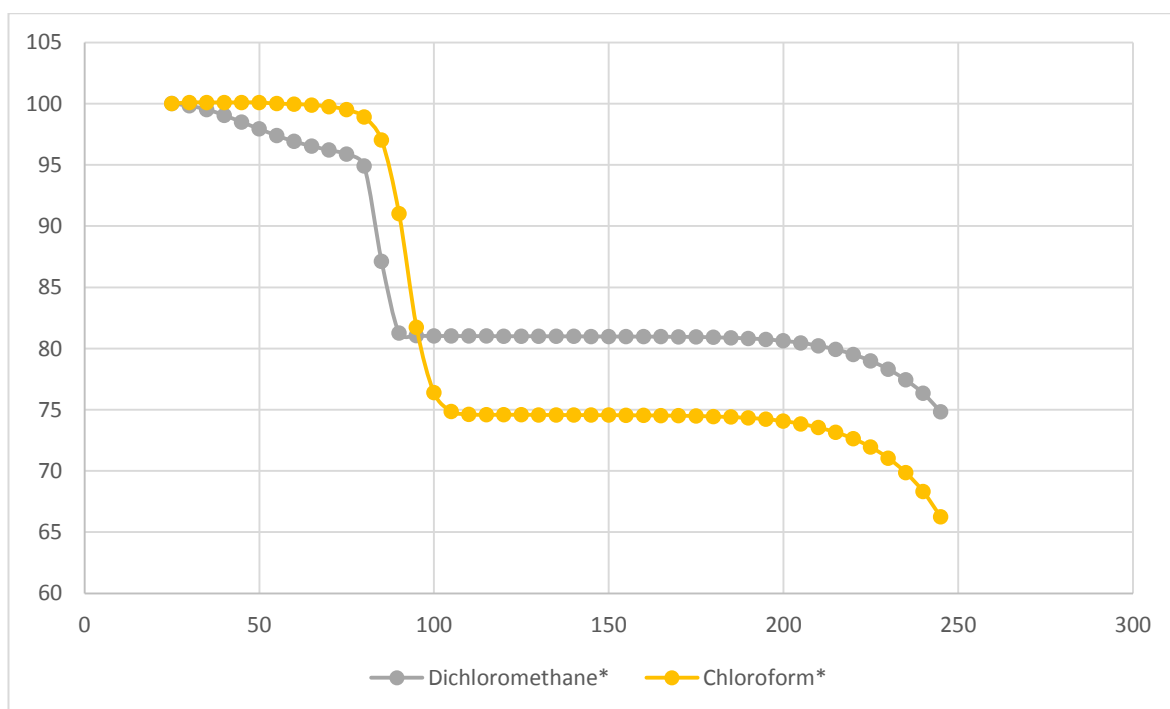


Figure 7-16. TGA profiles of the dichloromethane and chloroform solvate forms of griseofulvin.

The loss of total weight of the dichloromethane and chloroform solvate upon desolvation was (19 %) (25.4 %), respectively. Both solvents in a 1:1 solvate (the theoretical weight loss for 1:1 dichloromethane solvate is 19.4 % and for 1:1 chloroform solvate – 25.3 %). These results match the forms that were previously reported in the literature.^{56, 57} The crystal structure of the dichloromethane solvate had not been reported. For this reason, a single crystal X-ray diffraction experiment was conducted for this solvate. The results and discussion of the structure are presented in section 7.5.

7.4 Prediction vs results

After all screening experiments were completed; it was time to compare the experimental results with the predictions made via the computer. The predictions made by each model are presented separately in Tables 7-2 to 7-6. The highlighted cells in these tables indicate a wrong prediction by the models.

Table 7-2. Prediction vs. screening results for the ethanol predictive model

Drug candidate	Prediction value	Experimental solvate
Carbamazepine	0.427	No
Felodipine	0.623	No
Diflunisal	0.649	No
Griseofulvin	0.657	No
Ketoconazole	0.709	No
Theophylline	0.827	No
Hymecromone	0.866	No
Isoniazid	0.875	No
Fenofibrate	0.901	No
Ethenzamide	0.908	No

Table 7-3. Prediction vs. screening results for the methanol predictive model

Drug candidate	Prediction value	Experimental Solvate
Carbamazepine	0.455	No
Ketoconazole	0.570	No
Isoniazid	0.576	No
Diflunisal	0.601	No
Ethenzamide	0.715	No
Felodipine	0.736	No
Hymecromone	0.736	Yes
Theophylline	0.752	No
Griseofulvin	0.771	No
Fenofibrate	0.837	No

Table 7-4. Prediction vs. screening results for the dichloromethane predictive model

Drug candidate	Prediction value	Experimental Solvate
Ketoconazole	0.439	No
Griseofulvin	0.668	Yes
Felodipine	0.680	No
Carbamazepine	0.791	No
Fenofibrate	0.796	No
Diflunisal	0.824	Yes
Theophylline	0.929	No
Hymecromone	0.953	No
Ethenzamide	0.969	No
Isoniazid	0.970	No

Table 7-5. Prediction vs. screening results for the chloroform predictive model

Drug candidate	Prediction value	Experimental Solvate
Ketoconazole	0.472	No
Felodipine	0.667	No
Griseofulvin	0.684	Yes
Carbamazepine	0.770	No
Fenofibrate	0.790	No
Diflunisal	0.813	Yes
Theophylline	0.928	No
Hymecromone	0.943	No
Ethenzamide	0.965	No
Isoniazid	0.975	No

Table 7-6. Prediction vs. screening results for the water predictive model

Drug candidate	Prediction value	Experimental hydrate
Ketoconazole	0.079	No
Fenofibrate	0.243	No
Felodipine	0.291	No
Griseofulvin	0.304	No
Carbamazepine	0.320	Yes
Diflunisal	0.539	No
Hymecromone	0.658	Yes
Ethenzamide	0.701	No
Theophylline	0.710	Yes
Isoniazid	0.803	No

The ethanol model has performed unexpectedly well with 9 correct predictions out of 10. The only misprediction by this model was for the carbamazepine molecule. It is worth noting that no solvate was formed by the 10 candidates that were tested. This probably shows the ethanol model predictive ability towards non-solvate forming molecules only. The methanol model followed with a success rate of 8 out of 10. The misprediction was for one solvate (hymecromone) and one non-solvate (carbamazepine). The dichloromethane and chloroform models have shown a lower predictive ability than the previous models, with 3 mispredictions for each model. These wrong predictions were for one non-solvate (ketoconazole) and two solvate forms (griseofulvin and diflunisal). The water model has mispredicted the behaviour of 5 out of 10 compounds. Ketoconazole, felodipine and fenofirate were predicted to form a hydrate but didn't do so. The other two mispredictions, hmyecromone and theophylline were predicted not to form a hydrate, yet they were able to form one.

As the main reasons for solvate formation are not known, the real reasons for the mispredictions are not known either. An explanation could be established for each molecule

separately, in terms of molecular descriptors and 3D structure; however this would take us out of the scope of the research, where we are trying to find a trend for solvate formation. Alternatively, some patterns have been noticed in this set of tested molecules. For example, izonizaid and ethenzamide had a very small size, resulting in the formation of no solvate with any of the 5 solvents, despite their hydrogen bonding ability. Ketoconazole, fenofibrate and felodipine were large in size, yet failed to form a solvate with any of the solvent. It can be noticed that the first two possess no hydrogen bond donors, which can explain their behaviour not to form solvates. Felodipine does possess one hydrogen bond donor, but the ratio of hydrogen bond donors to its size is small. Moreover, all three molecules possessed a chlorinated ring, which could signal the existence of an electrostatic factor contributing to strong intermolecular interactions; therefore, solvent exclusion from the crystal.

For the molecules that were able to form a solvate despite the model predicting them not to form one, it seems like these molecules have the ability to arrange themselves in a low-energy solvate form. One trend that was obvious among these was special to the water model, where rigid molecules (consisting of connected rings with some degree of branching) were able to form a hydrate; these are hymecromone, theophylline anhydrous and carbamazepine.

7.5 Griseofulvin dichloromethane solvate

7.5.1 Overview

Griseofulvin formed a dichloromethane solvate, as determined from TG results shown in Figure 7-16. Further investigation regarding this solvate will be carried out in this section. The griseofulvin dichloromethane solvate was selected among all other solvates due to the interesting properties it showed. For example, colourless single crystals of the solvate were formed upon slurrying, yet, the crystal structure of this compound was not known. Additionally, the form was stable for more than 24 h at ambient conditions where no changes in TGA profiles were observed between a fresh and a stored sample.

7.5.2 Under the microscope

An optical refractive microscope and an optical transmittance microscope were used for observing the griseofulvin dichloromethane solvate crystals (see section 3.2.6 for more details about the microscopes). Both of these microscopes were equipped with heating units (hot-stage microscope). The heating units help studying the morphological changes of the crystals over a range of temperature, particularly around the desolvation temperature (80 °C). Since the temperature required to achieve the desolvation wasn't too high, the samples were heated under the microscope to 120 °C with a rate of 5 °C min⁻¹. The results are shown in Figure 7-17.

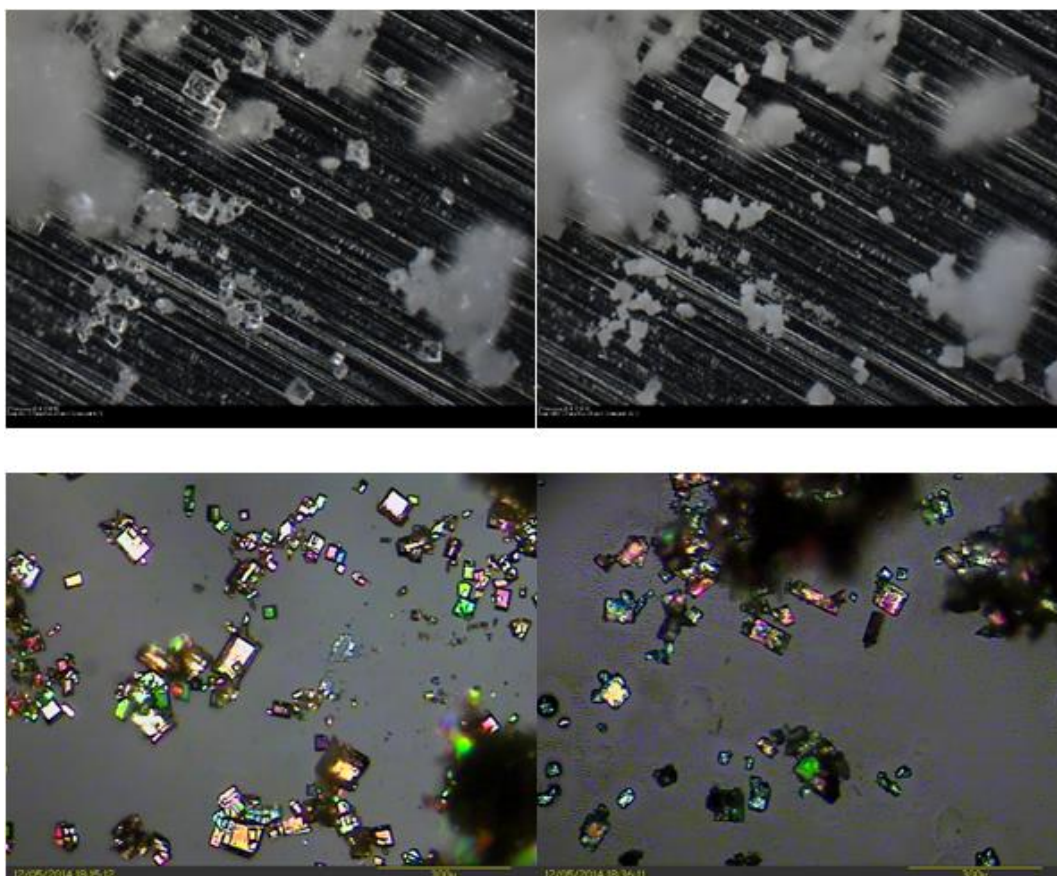


Figure 7-17. The first and the last frame of the heating cycles of griseofulvin dichloromethane solvate in the reflective (a) and transmittance microscope (b). The first images (left) were taken at ambient room temperature while the final images (right) were taken at 120 °C.

In both techniques, the griseofulvin started as transparent brick/slate-shaped crystals. Upon heating, the transparency of these crystals was lost while the morphology of the crystals was retained.

7.5.3 X-ray data and structure solution

A crystal of suitable size for single crystal X-ray diffraction analysis was obtained from the filtered and dried slurry. The cell parameters determined for this crystal in comparison to cell parameters for griseofulvin dichloromethane solvate available from literature are given in Table 7-7. The cell parameters reported previously were determined using a capillary method and calculated *via* least squares approach. The space group was not reported.⁵⁷

Table 7-7. Crystallographic parameters of the new dichloromethane solvate (II) versus the reported one (I)

<i>Parameter</i>	<i>New DCM solvate II</i>	<i>Reported DCM solvate I</i>
<i>Lattice system</i>	Monoclinic	Triclinic
<i>Space group</i>	<i>I</i> 2	Not reported
<i>a</i> (Å)	11.7585(6)	11.776(5)
<i>b</i> (Å)	8.5592(4)	11.918(6)
<i>c</i> (Å)	19.6721(13)	8.640(4)
α (°)	90.0	111.44(3)
β (°)	96.817(5)	90.00(3)
γ (°)	90.0	66.69(3)
<i>Cell volume</i>	1965.86(19)	-
<i>Density</i>	1.482	-
<i>Crystal size</i> (mm)	0.4 x 0.04 x 0.02	-
<i>Z</i>	2	-
<i>R</i>	0.0592	-
<i>wR2</i>	0.1542	-
<i>Temperature</i> (K)	293(2)	-

The cell parameter comparison of the solvate obtained in this work and that reported in literature shows considerable differences in the crystallographic data. The new solvate crystallized in the space group *I*2 (monoclinic) as opposed by the triclinic system reported in literature.⁵⁷ The crystal structure data acquired within this work were used to simulate a theoretical powder pattern of griseofulvin dichloromethane solvate. This simulated pattern was compared to the pattern acquired from literature in order to identify whether a new crystalline form has been obtained, as illustrated in Figure 7-18.

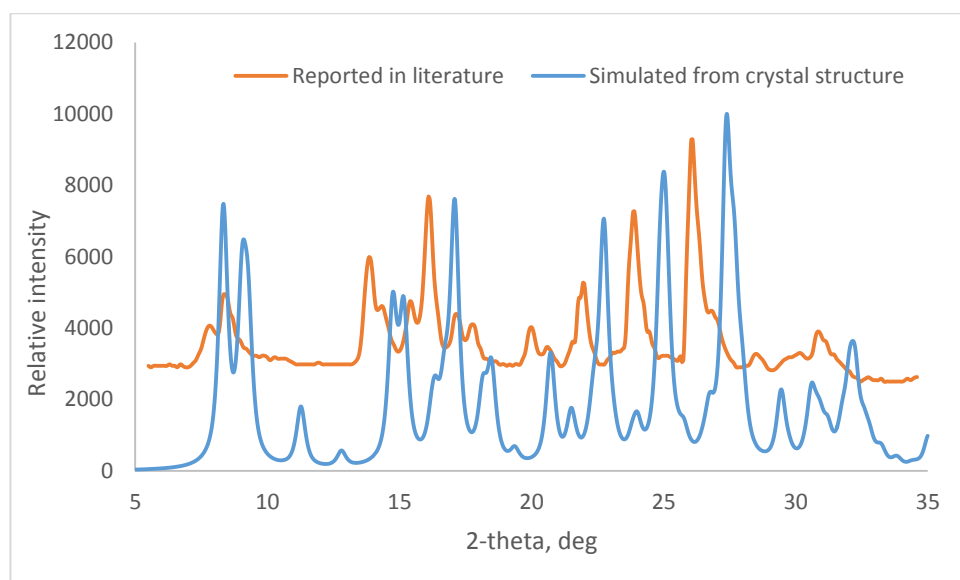


Figure 7-18. Comparison of X-ray diffraction patterns of griseofulvin dichloromethane solvate simulated from crystal structure (blue) and reported in literature (orange).

The comparison of the two X-ray diffraction patterns show some similarity although diffraction peak positions are considerably shifted probably due to temperature difference (the single crystal data were collected at lower temperature). The apparent similarity between both patterns imply that the same crystal form as reported in literature has been obtained in this work. However, since crystal structure of the solvate had not been determined previously, it is reported in this thesis. The molecular structure of this solvate is shown in Figure 7-19.

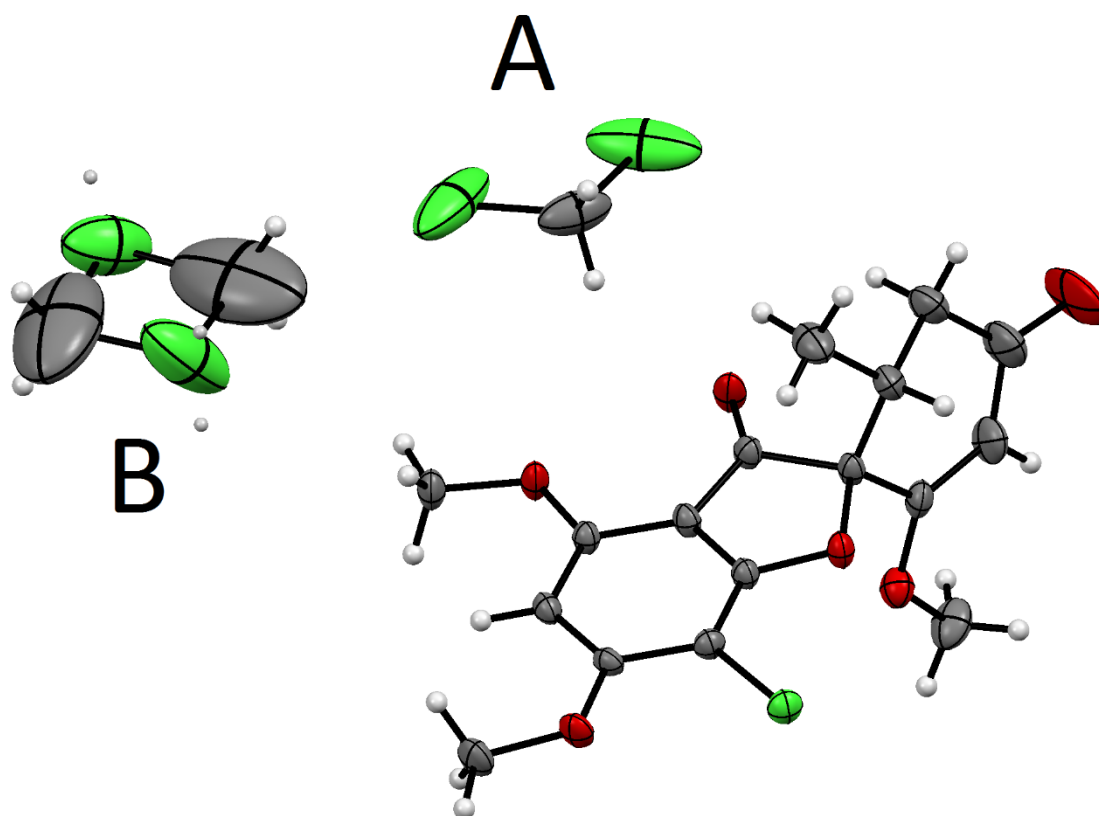


Figure 7-19. Molecular structure of griseofulvin showing ellipsoids of thermal displacement parameters. Probability of shown ellipsoids is set to 50 %.

The crystal structure solution of the griseofulvin dichloromethane solvate showed two crystallographically independent solvent molecules in the asymmetric unit. The two solvent molecules are situated on special positions (i.e. 2 fold symmetry axes), and therefore only half of each molecule belongs to the asymmetric unit. Moreover, one of the dichloromethane molecules shows a disorder over multiple positions. Several datasets were obtained from different crystals and all of them have shown the same trend of one ordered and one disordered dichloromethane molecule. The disorder in the solvent molecules has been accounted for by refining the occupancies of the corresponding atoms. The chlorine atoms of molecule A (see Figure 7-19) are symmetry-related and have occupancy of 1. The dichloromethane molecule B is disordered over multiple positions. As a result, each of the four atom positions can be occupied either by chlorine or by carbon. An illustration of this disorder, along with the atom numbering is shown in Figure 7-20.

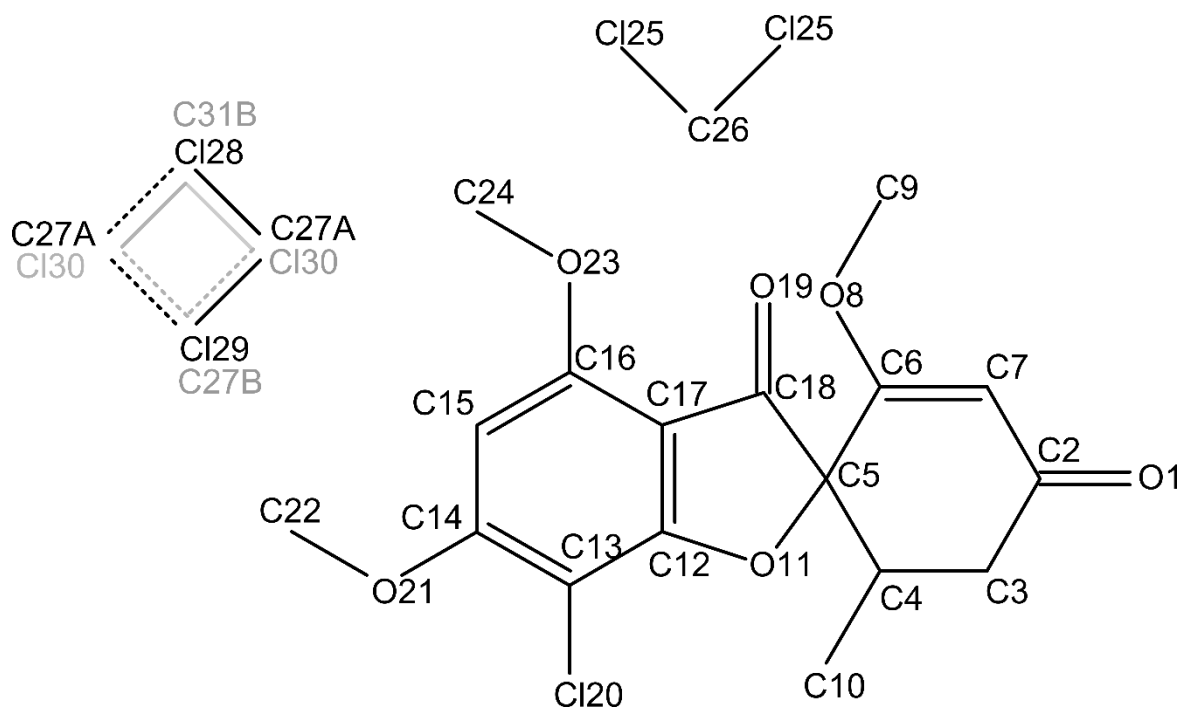


Figure 7-20. An illustration of the numbering of griseofulvin.

In the disorder component of the disordered dichloromethane (molecule B) illustrated by black lines, the occupancy of the carbon atoms (C27A) should be halved because the carbon atom can reside on one of the two symmetrically related positions shown in Figure 7-20 by either continuous or dashed lines. The chlorine atoms also should have an occupancy of 0.5, because they are on a twofold axis. When the grey illustration of the disordered dichloromethane molecule in Figure 7-20 is considered, the carbon atoms (C27B, C31B) represent two alternative orientations of the molecule. The proportion of former (black) to the latter (grey) part was 0.21:0.29, leading to a total occupancy of this disordered moiety of 0.5, per asymmetric unit. Together with the non-disordered dichloromethane molecule the total stoichiometry of the solvate is 1:1, as observed in the TGA thermogram in Figure 7-16.

Previous investigations have reported the crystal structures of griseofulvin polymorphs as well as a chloroform solvate. The overlays of the griseofulvin molecule in the newly found solvate with the reported griseofulvin structures^{56, 58-60} and with the known chloroform solvate⁵⁶ are shown in Figure 7-21 (a) and (b), respectively.

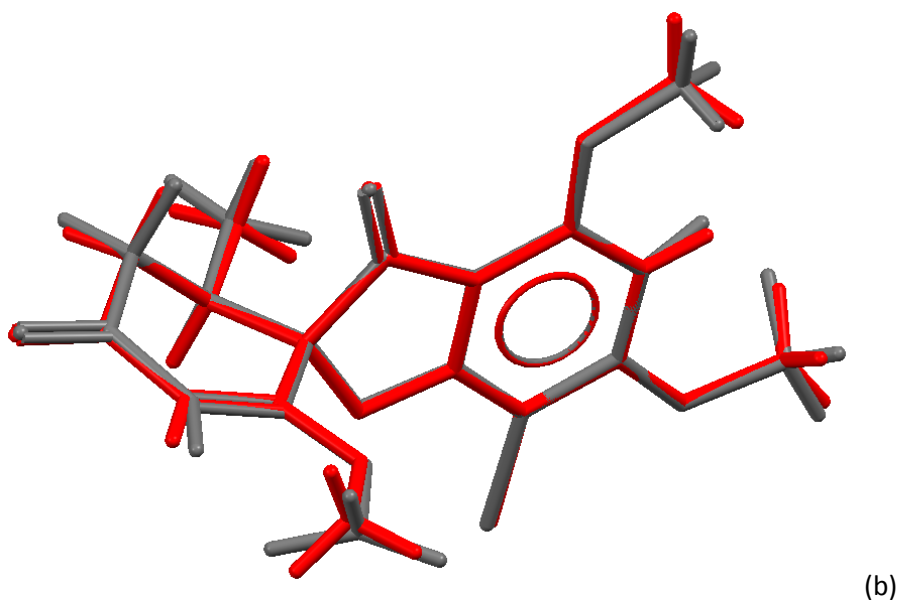
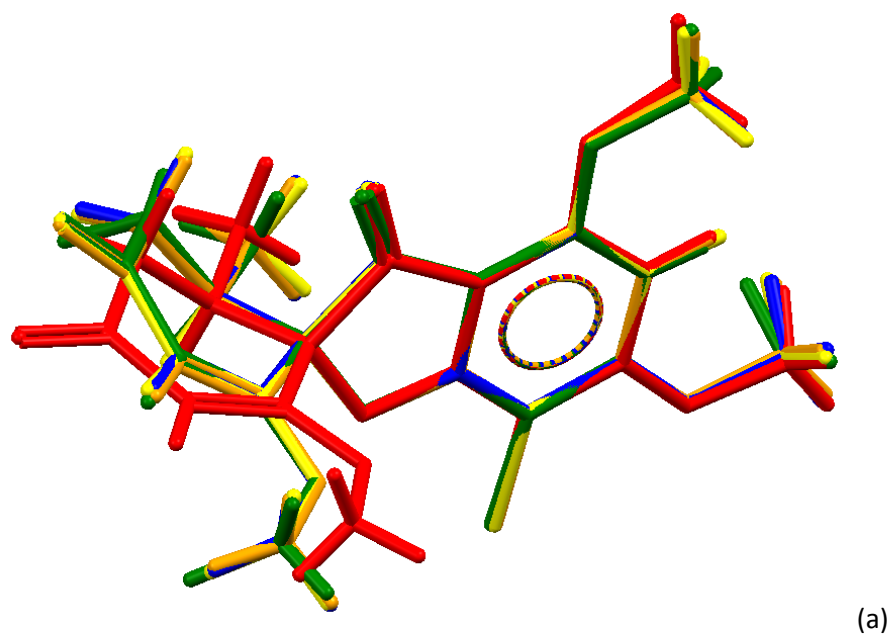


Figure 7-21. (a)An overlay of the griseofulvin molecule in the new DCM solvate (red) with all reported structures of griseofulvin in the CSD. Reference code GRISFL is shown in blue, GRISFL02 is shown in yellow, GRISFL03 is shown in orange, GRISFL04 is shown in green. (b) An overlay of the new griseofulvin dichloromethane solvate (red) with the exiting chloroform solvate (grey), CSD reference code MATZEO.

The overlay in part (a) of the figure shows that the solvate and all polymorphs exhibit a similar conformation in the rigid part of the structure (the fused rings and their branches), while they show a difference in the direction of the cyclohexane ring and the radicals attached to it. On the other hand, part (b) of Figure 7-21 shows high similarity between the dichloromethane and

chloroform solvate in terms of the molecular structure. Crystal packing similarity function in mercury was applied for the dichloromethane and chloroform solvate, the result was 9 out of 15 molecules in common.

7.5.4 Interactions and packing

Before discussing the structure, it is useful to note that in this section, as in chapter 6, any interactions with inter-atomic distances shorter than the VDW radii (of interacting atoms) by at least 0.1 Å are referred to as “short interactions”. It is also important to note that the numbering that is shown in Figure 7-20 is going to be used for the discussion of the interactions and packing.

Looking at the interactions of griseofulvin molecules, it can be seen that each molecule forms a dimer with an inverted molecule *via* a pair of C–H \cdots π interactions. Specifically, the interaction takes place between H10A and C15 on the chlorinated benzene ring, [$d(\text{H}\cdots\text{C})=2.787$ Å, $\angle\text{C}-\text{H}\cdots\text{C}=147.64^\circ$] as illustrated in Figure 7-22(a). Note the high contact surface area between the molecules in the dimer associated with this interaction, which enforces the connection of the molecules in this dimer further. These dimers are connected to other dimers along the crystallographic *a* axis *via* C–H \cdots O interactions resulting in a ladder-like arrangement, as shown in Figure 7-22(a). These weak hydrogen bonds take place between O1 and H24A [$d(\text{H}\cdots\text{O})=2.454$ Å, $\angle\text{C}-\text{H}\cdots\text{O}=123^\circ$]. The ladder-like motif is connected to other similar motifs in horizontal and diagonal directions if the “packing” was viewed along the *a* axis, as illustrated Figure 7-22(b). Two interactions linking the dimer motif to another horizontally were noticed, these are a short halogen interaction between Cl20 and O23 [$d(\text{H}\cdots\text{O})=3.137$ Å, $\angle\text{C}-\text{H}\cdots\text{O}=140.10^\circ$] in addition to a slightly longer C–H \cdots O bond between H24C and O21 [$d(\text{H}\cdots\text{O})=2.669$ Å, $\angle\text{C}-\text{H}\cdots\text{O}=154.18^\circ$]. These interactions are viewed alone in Figure 7-22(c). Diagonally, the dimers are held *via* a two C–H \cdots O bonds between H9C and O19 [$d(\text{H}\cdots\text{O})=2.424$ Å, $\angle\text{C}-\text{H}\cdots\text{O}=164.37^\circ$] and O11-H9A [$d(\text{H}\cdots\text{O})=2.734$ Å, $\angle\text{C}-\text{H}\cdots\text{O}=150.65^\circ$]. Note

that the H9C and H9A mentioned here belong to two different molecules. These two bonds result in a ring-like $R_2^2(13)^{61}$ hydrogen bonded dimer, as illustrated in Figure 7-22(d).

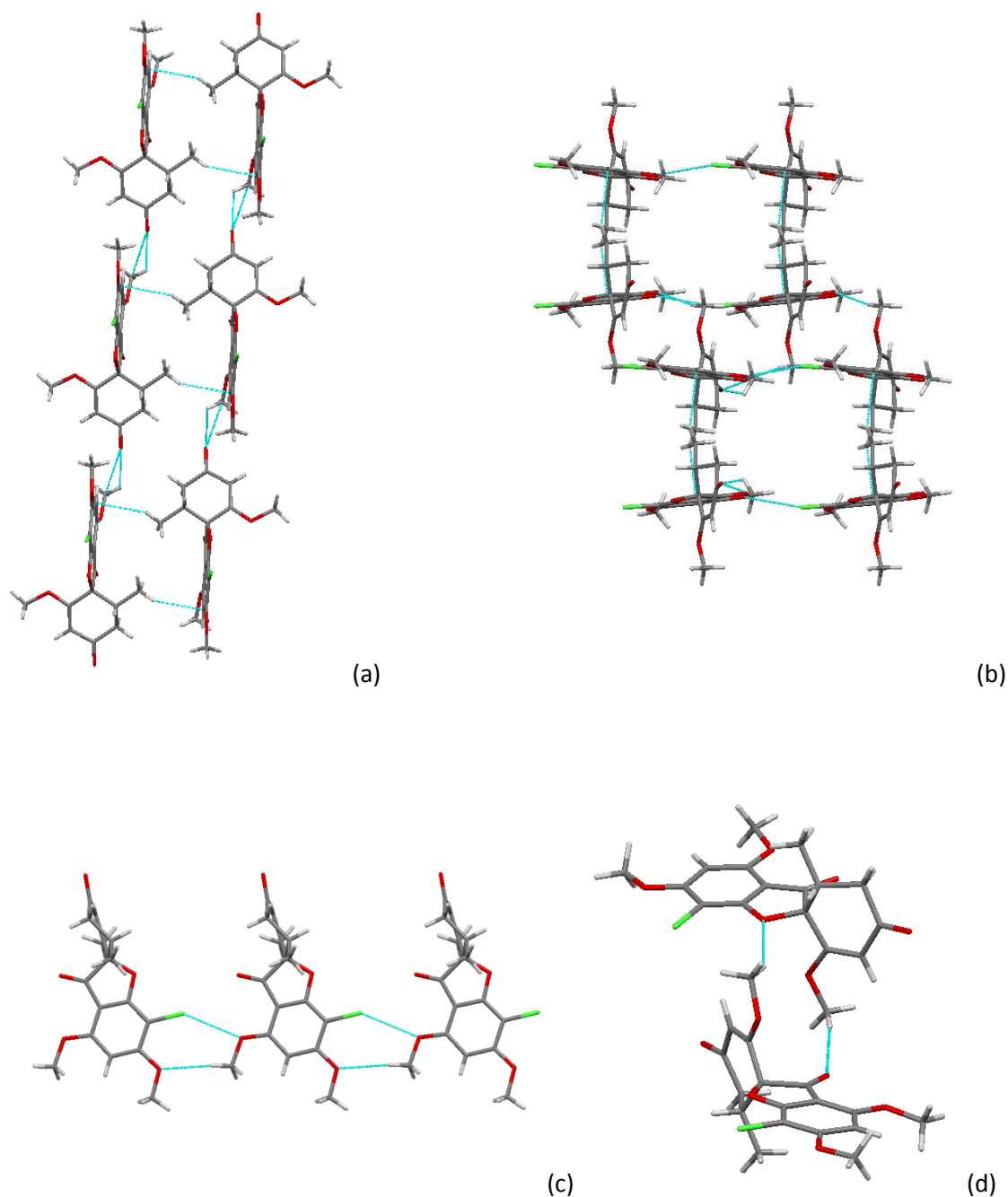


Figure 7-22. The interactions between griseofulvin molecules. Part (a) shows the dimers formed by griseofulvin molecules along the B axis. Part (b) shows four pairs of dimers along the a axis. Part (c) shows the horizontal interaction between the molecules in part (b), along the C axis. Part (d) shows the ring-like interaction that connects the dimers in part (b) diagonally.

The interactions between the griseofulvin molecules result in a structure with channels, where the dichloromethane molecules are positioned. These channels are illustrated in Figure 7-23.

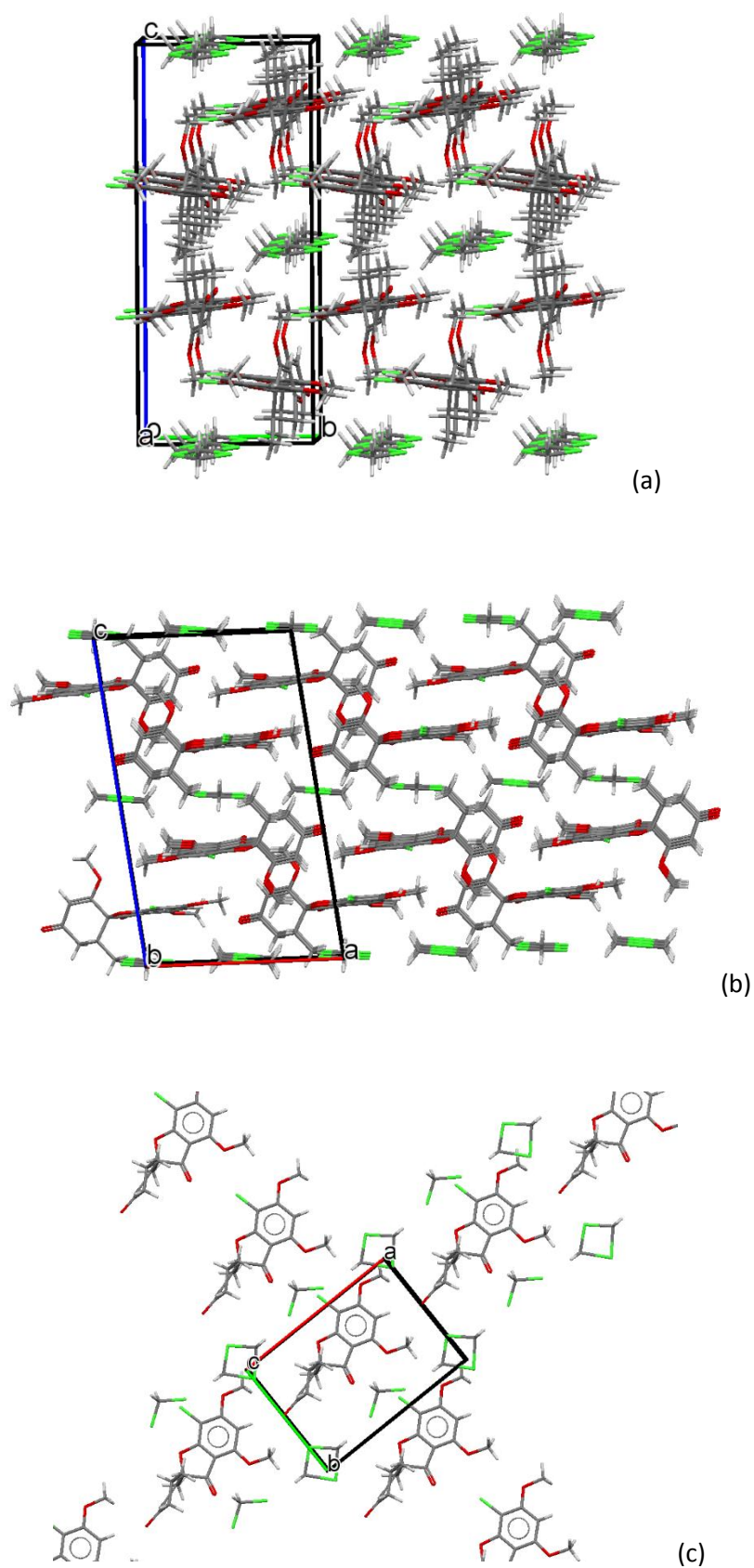


Figure 7-23. The channels of dichloromethane in the griseofulvin dichloromethane solvate II from. Part (a) shows these channels along the a axis, part (b) show the same channels along the b axis and part (c) show the same channels along the c axis.

It can be noticed that each channel contains a sequence of disordered-ordered dichloromethane molecules as can be seen in Figure 7-23(c). The non-disordered dichloromethane molecules have short interactions with two griseofulvin molecules. These are formed through two symmetry-equivalent C–H \cdots O bonds. Specifically, it is between the hydrogen atoms in the dichloromethane molecule (H26A, H26B) and the O19 of two griseofulvin molecules [$d(\text{H}\cdots\text{O}) = 2.586 \text{ \AA}$, $\angle\text{C–H}\cdots\text{O} = 142^\circ$]. This dichloromethane molecule also forms short interactions with adjacent, disordered dichloromethane molecules through a C–H \cdots Cl bond ($d(\text{H}\cdots\text{O}) = 2.52 \text{ \AA}$, $\angle\text{C–H}\cdots\text{O} = 121^\circ$). With exception of the latter weak bond, the disordered dichloromethane molecules were not involved in any other short attractive interaction, allowing free rotation of these molecules, resulting in the disordered fragment.

7.6 References

1. Barnes PJ. Theophylline. American Journal of Respiratory and Critical Care Medicine. 2013;188(8):901-6.
2. ZuWallack RL, Mahler DA, Reilly D, Church N, Emmett A, Rickard K, et al. Salmeterol Plus Theophylline Combination Therapy in the Treatment of COPD. CHEST Journal. 2001;119(6):1661-70.
3. Hansel TT, Tennant RC, Tan AJ, Higgins LA, Neighbour H, Erin EM, et al. Theophylline: Mechanism of action and use in asthma and chronic obstructive pulmonary disease. Drugs of Today. 2004;40(1):55-69.
4. Barnes PJ, Pauwels RA. Theophylline in the management of asthma: time for reappraisal? European Respiratory Journal. 1994;7:579-91.
5. Ward AJM, McKenniff M, Evans JM, Page CP, Costello JF. Theophylline—an Immunomodulatory Role in Asthma? American Review of Respiratory Disease. 1993;147(3):518-23.
6. Abate A, Dimartino V, Spina P, Costa PL, Lombardo C, Santini A, et al. Hymecromone in the treatment of motor disorders of the bile ducts: a multicenter, double-blind, placebo-controlled clinical study. Drugs under Experimental and Clinical Research. 2001;27(5-6):223-31.
7. Nagy N, Kuipers HF, Frymoyer AR, Ishak HD, Bollyky JB, Wight TN, et al. 4-Methylumbelliferone treatment and hyaluronan inhibition as a therapeutic strategy in inflammation, autoimmunity, and cancer. Frontiers in immunology. 2015;6:123.
8. Yates TJ, Lopez LE, Lokeshwar SD, Ortiz N, Kallifatidis G, Jordan A, et al. Dietary Supplement 4-Methylumbelliferone: An Effective Chemopreventive and Therapeutic Agent for Prostate Cancer. Journal of the National Cancer Institute. 2015;107(7) djv085.

9. van den Boogaard J, Kibiki GS, Kisanga ER, Boeree MJ, Aarnoutse RE. New drugs against Tuberculosis: Problems, Progress, and evaluation of Agents in Clinical Development. *Antimicrobial Agents and Chemotherapy*. 2009;53(3):849-62.
10. World Health Organization. Global tuberculosis control: WHO report 2010
11. Dickinson JM, Ellard GA, Mitchison DA. Suitability of isoniazid and ethambutol for intermittent administration in the treatment of tuberculosis. *Tubercle*. 1968;49(4):351-66.
12. Darias V, Bravo L, Abdallah SS, Sanchez Mateo CC, Exposito-Orta MA, Lissavetsky J, et al. Synthesis and Preliminary Pharmacological Study of Thiophene Analogues of the Antipyretic and Analgesic Agent Ethenzamide. *Archiv der Pharmazie (Weinheim)*. 1992;325(2):83-7.
13. Kawano O, Sawabe T, Misaki N, Fukawa K. Studies on Combination Dosing (III) Aspirin and Ethenzamide. *The Japanese Journal of Pharmacology*. 1978;28(6):829-35.
14. Yasuno N, Tsuchiya M, Kizu J, Watanabe M, Arakawa Y, Umeyama T, et al. Development of Ethenzamide Ointment as a Pain Relief for Postherpetic Neuralgia. *Iryo Yakugaku (Japanese Journal of Pharmaceutical Health Care and Sciences)*. 2002;28(4):309-14.
15. Dalby MA. Antiepileptic and Psychotropic Effect of Carbamazepine (Tegretol®) in the Treatment of Psychomotor Epilepsy. *Epilepsia*. 1971;12(4):325-34.
16. Bird CAK, Griffin BP, Miklaszewska JM, Galbraith AW. Tegretol (Carbamazepine): A Controlled Trial of a New Anti-Convulsant. *The British Journal of Psychiatry*. 1966;112(488):737-42.
17. Campbell FG, Graham JG, Zilkha KJ. Clinical trial of carbazepine (Tegretol) in trigeminal neuralgia. *Journal of Neurology, Neurosurgery, and Psychiatry*. 1966;29(3):265-7.
18. Ballenger JC, Post RM. Carbamazepine in manic-depressive illness: a new treatment. *The American Journal of Psychiatry*. 1980;137(7):782-90.

19. Okuma T, Kishimoto A, Hisashi M, Atsushi O, Toji M, Nakao T, et al. Anti-Manic and Prophylactic Effects of Carbamazepine (Tegretol) on Manic Depressive Psychosis A Preliminary Report. *Psychiatry and Clinical Neurosciences*. 1973;27(4):283-97.
20. Post RM, Uhde TW, Ballenger JC, Squillace KM. Prophylactic efficacy of carbamazepine in manic-depressive illness. *The American Journal of Psychiatry*. 1983;140(12):1602-4.
21. Rull JA, Quibrera R, González-Millán H, Castañeda OL. Symptomatic treatment of peripheral diabetic neuropathy with carbamazepine (Tegretol®): double blind crossover trial. *Diabetologia*. 1969;5(4):215-8.
22. Lipper S, Davidson JR, Grady TA, Edinger JD, Hammett EB, Mahorney SL, et al. Preliminary study of carbamazepine in post-traumatic stress disorder. *Psychosomatics*. 1986;27(12):849-54.
23. Forbes JA, Beaver WT, White EH, White RW, Neilson GB, Shackleford RW. Diflunisal. A New Oral Analgesic With an Unusually Long Duration of Action. *The Journal of the American Medical Association*. 1982;248(17):2139-42.
24. Forbes JA, Calderazzo JP, Bowser MW, Foor VM, Shackleford RW, Beaver WT. A 12-Hour Evaluation of the Analgesic Efficacy of Diflunisal, Aspirin, and Placebo in Postoperative Dental Pain. *The Journal of Clinical Pharmacology*. 1982;22(2-3):89-96.
25. Sisk AL, Mosley RO, Martin RP. Comparison of preoperative and postoperative diflunisal for suppression of postoperative pain. *Journal of Oral and Maxillofacial Surgery*. 1989;47(5):464-8.
26. Forbes JA, Kolodny AL, Beaver WT, Shackleford RW, Scarlett VR. A 12-hour evaluation of the analgesic efficacy of diflunisal, acetaminophen, and acetaminophen-codeine combination, and placebo in postoperative pain. *Pharmacotherapy*. 1983;3(2 Pt 2):47s-54s.

27. Ellen RL, McPherson R. Long-Term Efficacy and Safety of Fenofibrate and a Statin in the Treatment of Combined Hyperlipidemia. *The American journal of cardiology*. 1998;81(4A):60B-5B.
28. Effect of fenofibrate on progression of coronary-artery disease in type 2 diabetes: the Diabetes Atherosclerosis Intervention Study, a randomised study. *The Lancet*. 2001;357(9260):905-10.
29. Keech AC, Mitchell P, Summanen PA, O'Day J, Davis TM, Moffitt MS, et al. Effect of fenofibrate on the need for laser treatment for diabetic retinopathy (FIELD study): a randomised controlled trial. *Lancet*. 2007;370(9600):1687-97.
30. Elmfeldt D, Hedner T. Antihypertensive effects of felodipine compared with placebo. *Drugs*. 1985;29(2):109-16.
31. Dahlöf B, Hosie J. Antihypertensive efficacy and tolerability of a new once-daily felodipine-metoprolol combination compared with each component alone. The Swedish/UK Study Group. *Blood Pressure Supplement*. 1993;1:22-9.
32. Heeres J, Backx LJJ, Mostmans JH, Van Cutsem J. Antimycotic imidazoles. Part 4. Synthesis and antifungal activity of ketoconazole, a new potent orally active broad-spectrum antifungal agent. *Journal of Medicinal Chemistry*. 1979;22(8):1003-5.
33. Van Cutsem J. The antifungal activity of ketoconazole. *The American Journal of Medicine*. 1983;74(1, Part 2):9-15.
34. Grycová A, Dořičáková A, Dvořák Z. Impurities contained in antifungal drug ketoconazole are potent activators of human aryl hydrocarbon receptor. *Toxicology Letters*. 2015;239(2):67-72.
35. Pont A, Williams PL, Loose DS, Feldman D, Reitz RE, Bochra C, et al. Ketoconazole Blocks Adrenal Steroid Synthesis. *Annals of Internal Medicine*. 1982;97(3):370-2.

36. Blank H, Roth FJ, BLANK H, ROTH FJ, Bruce WW, Engel MF, Smith JG, Zaias N. The treatment of dermatomycoses with orally administered griseofulvin. *AMA archives of dermatology*. 1959 Mar 1;79(3):259-66.
37. Lpez-Gmez S, Del Palacio A, Van Cutsem J, Soledad Cuetara M, Iglesias L, Rodriguez-Noriega A. Itraconazole versus griseofulvin in the treatment of tinea capitis: a double-blind randomized study in children. *International Journal of Dermatology*. 1994;33(10):743-7.
38. Gull K, Trinci APJ. Griseofulvin inhibits Fungal Mitosis. *Nature*. 1973;244(5414):292-4.
39. Liu K, Yan J, Sachar M, Zhang X, Guan M, Xie W, et al. A metabolomic perspective of griseofulvin-induced liver injury in mice. *Biochemical Pharmacology*. 2015;98(3):493-501.
40. Zhong N, Chen H, Zhao Q, Wang H, Yu X, Eaves AM, et al. Effects of griseofulvin on apoptosis through caspase-3- and caspase-9-dependent pathways in K562 leukemia cells: An in vitro study. *Current Therapeutic Research*. 2010;71(6):384-97.
41. Rebacz B, Larsen TO, Clausen MH, Rønneest MH, Löffler H, Ho AD, et al. Identification of Griseofulvin as an Inhibitor of Centrosomal Clustering in a Phenotype-Based Screen. *Cancer Research*. 2007;67(13):6342-50.
42. Bruns S, Reichelt J, Cammenga HK. Thermochemical investigation of theophylline, theophylline hydrate and their aqueous solutions. *Thermochimica Acta*. 1984;72(1):31-40.
43. Thareja S, Verma A, Kalra A, Gosain S, Rewatkar PV, Kokil GR. Novel chromeneimidazole derivatives as antifungal compounds: synthesis and in vitro evaluation. *Acta Poloniae Pharmaceutica-Drug Research*. 2010;67(4):423-7.
44. Good DJ, Rodríguez-Hornedo N. Solubility advantage of pharmaceutical cocrystals. *Crystal Growth and Design*. 2009;9(5):2252-64.
45. Brewer GA. Isoniazid. *Analytical Profiles of Drug Substances*. 1977;6:183-258.

46. Isoniazid. *Tuberculosis*.88(2):112-6.
47. THE JAPANESE PHARMACOPOEIA FOURTEENTH EDITION, PART I, Official monographs.
48. Martinez-Ohárriz MC, Martín C, Goni MM, Rodríguez-Espinoza C, Tros de Ilarduya-Apaolaza MC, Sánchez M. Polymorphism of diflunisal: isolation and solid-state characteristics of a new crystal form. *Journal of Pharmaceutical Sciences*. 1994;83(2):174-7.
49. Tipduangta P, Takieddin K, Fábán L, Belton P, Qi S. A New Low Melting-Point Polymorph of Fenofibrate Prepared via Talc Induced Heterogeneous Nucleation. *Crystal Growth & Design*. 2015;15(10):5011-20.
50. Surov AO, Solanko KA, Bond AD, Perlovich GL, Bauer-Brandl A. Crystallization and Polymorphism of Felodipine. *Crystal Growth & Design*. 2012;12(8):4022-30.
51. Van den Mooter G, Wuyts M, Blaton N, Busson R, Grobet P, Augustijns P, et al. Physical stabilisation of amorphous ketoconazole in solid dispersions with polyvinylpyrrolidone K25. *European Journal of Pharmaceutical Sciences*. 2001;12(3):261-9.
52. Yang D, Kulkarni R, Behme RJ, Kotiyan PN. Effect of the melt granulation technique on the dissolution characteristics of griseofulvin. *International Journal of Pharmaceutics*. 2007;329(1–2):72-80.
53. Sutor DJ. The structures of the pyrimidines and purines. VI. The crystal structure of theophylline. *Acta Crystallographica*. 1958;11(2):83-7.
54. Jasinski JP, Woudenberg RC. 7-Hydroxy-4-methylcoumarin monohydrate. *Acta Crystallographica Section C*. 1994;50(12):1952-3.
55. Reck G, Dietz G. The Order-Disorder Structure of Carbamazepine Dihydrate: 5 H-Dibenz [b, f] azepine-5-carboxamide Dihydrate, C₁₅H₁₂N₂O · 2 H₂O. *Crystal Research and Technology*. 1986;21(11):1463-8.

56. Cheng KC, Shefter E, Srikrishnan T. Crystal structure analysis of the desolvation of the chloroform solvate of griseofulvin. *International Journal of Pharmaceutics*. 1979;2(2):81-9.
57. Shirotani K-I, Suzuki E, Morita Y, Sekiguchi K. Solvate Formation of Griseofulvin with Alkyl Halide and Alkyl Dihalides. *Chemical & Pharmaceutical Bulletin*. 1988;36(10):4045-54.
58. Malmros G, Wagner A, Maron L. (2S,6'R)-7-chloro-2',4,6,-trimethoxy-6'-methyl-spiro-(benzofuran-2(3H),2-(2')cyclohexene)-3,4'-dione C₁₇H₁₇ClO₆. CRYSTAL STRUCTURE COMMUNICATIONS. 1977;6:463-70.
59. Puttaraja, Nirmala KA, Sakegowda DS, Duax WL. Crystal structure of griseofulvin. *Journal of Crystallographic and Spectroscopic Research*. 1982;12(5):415-23.
60. Loew E., Steglich W., Polborn K. CSD private communication. 2004.
61. Etter MC. Encoding and decoding hydrogen-bond patterns of organic compounds. *Accounts of Chemical Research*. 1990;23(4):120-6.

Chapter 8: Single crystal analysis of the new fenofibrate forms

8.1 Overview

This chapter is dedicated to describing new advances related to the polymorphs of the fenofibrate drug. The structural formula of fenofibrate and atom numbering used further is shown in Figure 8-1.

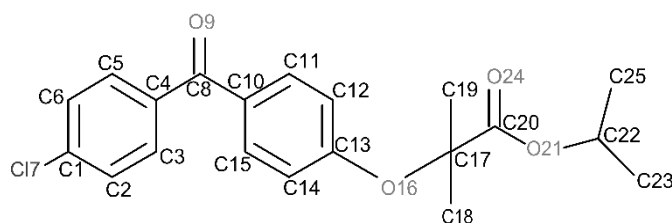


Figure 8-1. The structural formula of fenofibrate showing atom numbering scheme.

This anti-hyperlipdemic agent shows interesting crystallization properties when crystallized from its amorphous form. Homogeneous nucleation and crystallization are slow for amorphous fenofibrate.¹ Alternatively, heterogeneous crystallization can be initiated by addition of an impurity or mechanical stimulation.² The crystallization of fenofibrate was studied with a colleague working in the same suite, Mr Pratchaya Tipduangta. In this chapter, the theme of Tipduangta's work on fenofibrate will be outlined to give a background, but the main focus will be the part which I took part in, that is, the determination of the crystal structure of two fenofibrate polymorphs, form IIa and form III using single crystal XRD.

8.2 Fenofibrate polymorphs

Forms I and II of fenofibrate were reported in the literature.³⁻⁶ The literature regarding form I was clear, where multiple papers described the same crystalline form.³⁻⁵ On the other hand, the literature about form II was not as consistent. Di Martino *et al.* were the first to report form II of fenofibrate in 2000, where they provided PXRD pattern and DSC data of this form. More details about the same form was given by Heinz *et al.* in 2009.^{3, 4} The latter has

confirmed obtaining the same form as Di Martino's group using DSC data. Another group, Balendiran *et al.* have reported a crystal structure of form II in 2011.⁵ What is surprising is that the simulated PXRD pattern of fenofibrate in the latter work does not match the recorded PXRD pattern obtained by the former groups, which signals these are different forms. To avoid confusion, Tipduangta has denoted the form discovered by Di Martino *et al.* as "form IIa" while he denoted the form reported by Balendiran *et al.* as "form IIb". The PXRD of these forms are shown in Figure 8-2.

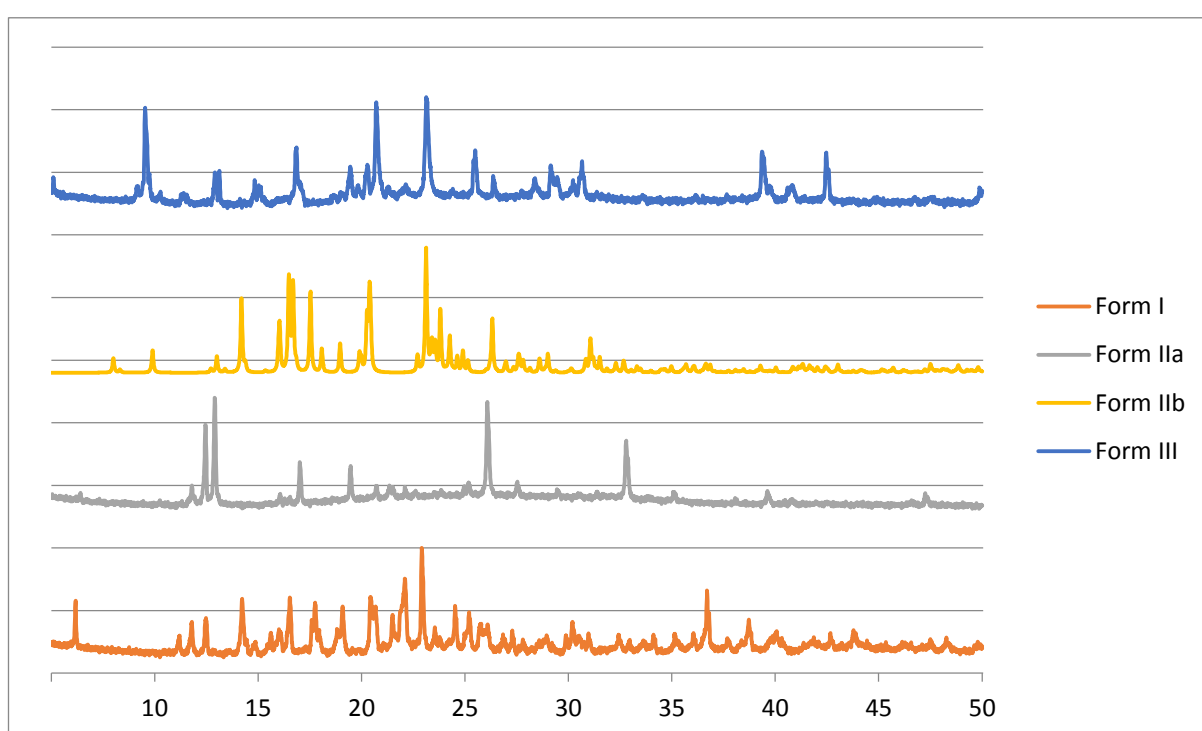


Figure 8-2. PXRD patterns of the different fenofibrate polymorphs.

8.3 Polymorph preparation

The method that was used to obtain each of the four fenofibrate forms is described in this section. Form I of fenofibrate is simply the powder that was obtained from the supplier (Sigma-Aldrich). Form IIa of fenofibrate was obtained *via* heterogeneous crystallization of amorphous fenofibrate. The preparation started with placing the fenofibrate powder (form I) on a microscopic slide, where it was heated to a 100 °C until all fenofibrate was melted. After

that, the solid on the microscopic slide was left to cool at room temperature to obtain amorphous material. Crystallization was initiated by surface disruption; that is, scratching the surface of the amorphous film with a stainless steel spatula. The microscopic slide was left to crystallize at room temperature, where it yields a mixture of fenofibrate forms I and IIa, with form I being dominant. Tipduangta manipulated two factors to get preferential crystallization of form IIa.⁸ The factors changed were the temperature and the free surface available during crystallization. The result was the possibility to obtain a pure form IIa by reducing the free surface available, i.e. by placing a coverslip on the microscopic slide right after the mechanical stimulation of the amorphous film and allowing crystallization to happen in the range from room temperature to 40 °C. It was also possible to obtain a mixture of fenofibrate form I and IIa, with form IIa being the dominant form by annealing the free-surface microscopic slide at 40-50 °C.

Form IIb was reported by Balnderian *et al.*, where they were able to obtain that form by recrystallization of fenofibrate from ethanol at room temperature. However, Tipduangta was not able to reproduce the form when the same procedure was followed. Alternatively, form I of fenofibrate was obtained. A group that worked on fenofibrate in 2014 had the same result as Tipduangta,⁷ where they obtained fenofibrate form I by crystallizing it from an undersaturated ethanol solution by slow evaporation.

Form III was obtained *via* heterogeneous nucleation of amorphous fenofibrate. In order to obtain this form, fenofibrate powder (form I) and talc (a pharmaceutical excipient) were weighed out in a ratio of 99:1, placed in a glass vial and mixed well with a stainless steel spatula. The vial was then heated to 100 °C until all fenofibrate melted. After that, it was transferred to a 0 % relative humidity desiccator. After 24 hours, the amorphous fenofibrate in the vial crystallized as form III. Obtaining this form was confirmed using microscopy, ATR-FTIR and PXRD.⁸

8.4 Single crystal X-ray Diffraction

8.4.1 Structure solution

The crystal structures of the forms IIa and III have not been previously determined. In order to obtain the crystal structure of these forms, single crystal X-ray diffraction experiments were conducted. The crystallographic parameters for these two forms compared to the forms mentioned in the literature are shown in Table 8-1.

Table 8-1. Crystallographic parameters of fenofibrate polymorphs

Parameter	Form I ⁹	Form IIa ⁸	Form IIb ⁵	Form III
Lattice system	Triclinic	Triclinic	Monoclinic	Triclinic
Space group	$P\bar{1}$	$P\bar{1}$	$P2_1/n$	$P\bar{1}$
a (Å)	8.1325	8.1328(5)	13.619	9.4803(6)
b (Å)	8.2391	8.7088(6)	7.554	9.7605(6)
c (Å)	14.399	13.6692(9)	17.88	10.9327(8)
α (°)	93.978	85.976(6)	90	110.840(6)
β (°)	105.748	84.815(5)	92.35	90.352(5)
γ (°)	95.854	74.344(6)	90	99.701(5)
Cell volume	919.03	927.34(11)	1837.909	929.53(11)
Density (gcm^{-3})	1.285	1.292	1.304	1.289(2)
Crystal size (mm)	0.55 x 0.50 x 0.44	0.12 x 0.13 x 0.46	0.55 x 0.30 x 0.25	0.12 x 0.15 x 0.34
Z	2	2	4	2
R	0.0418	0.0694	0.0355	0.0653
wR^2	0.105	0.1265	0.0897	0.149
Temperature (K)	193	140(2)	100	140
Goodness of fit	1.035	1.023	1.026	1.016

Both forms IIa and III crystallize in the triclinic $P\bar{1}$ space group with one molecule of fenofibrate in the asymmetric unit and two molecules in the unit cell. Structures of forms IIa and III have been solved according to the method described in section 8.3. The molecular structures of fenofibrate in these new forms are shown in Figure 8-3(a) and (b).

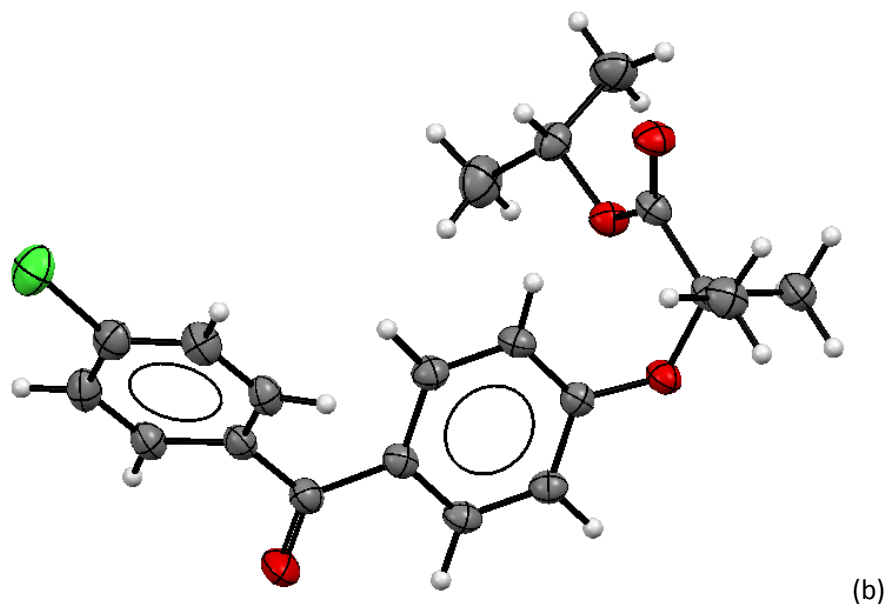
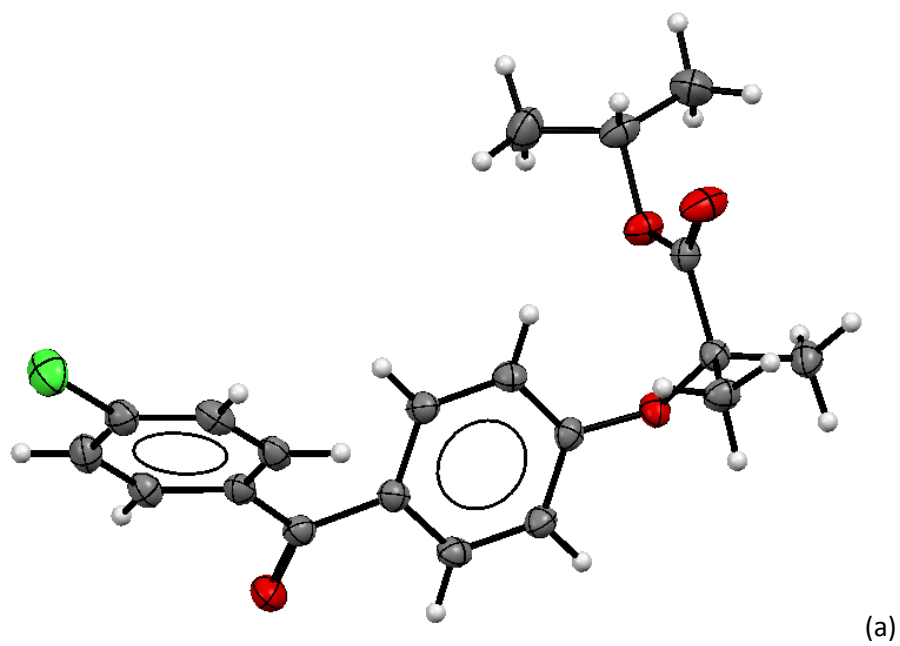


Figure 8-3. ORTEP structure of fenofibrate forms IIa (a) and III (b).

In order to visually compare the conformation of fenofibrate in its polymorphic forms, an overlay of the molecular structures in these forms was generated and is shown in Figure 8-4.

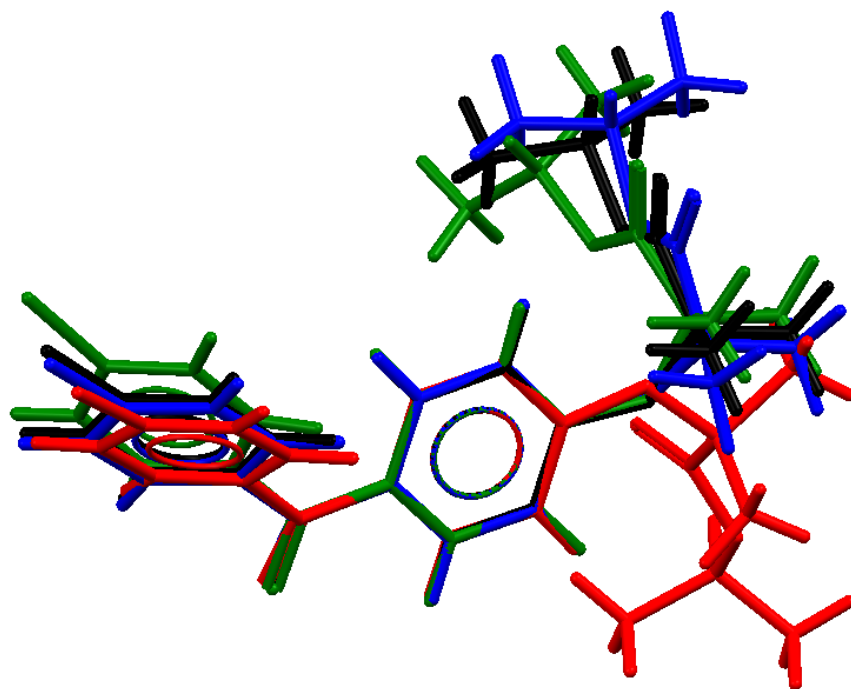


Figure 8-4. An overlay of fenofibrate form I (black), form IIa (blue), form IIb (red), form III (green).

The fenofibrate molecule can be divided into two main parts: a rigid part consisting of the two aromatic rings linked by a keto group and a flexible aliphatic tail. The overlay in Figure 8-4 shows that form IIa exhibits the highest proximity in spatial conformation to the most stable form I, followed by form III then form IIb. It is noticeable that for the 4 forms shown in the figure, the aliphatic chain is not fully extended, but rather folded back towards the aromatic part of the structure. This folding seems to be stabilized by an intramolecular C–H···O interaction between the middle benzene ring and the ester group in the alkyl chain. An overview of the intra-molecular interactions and angles representing the conformation of the molecule is given in Table 8-2.

Table 8-2. Intramolecular angles and short contact distances in fenofibrate. The RMSD is a function in mercury program that compares structural similarity between molecules/forms

Form	Angles between the two aromatic rings	Distance between carbonyl O of the ester and H of the central benzene ring	Distance between ether O and H of the benzene ring	Torsion angle (C14-C13-O16-C17)	RMSD compared to form I
I	48.62(7) °	2.81 (O24...H14)	2.84 (O21...H14)	-23.36	0
IIa	48.25(10) °	3.04 (O24...H14)	2.649(16) (O21...H14)	-40.35	0.189
IIb	53.73 °	2.90 (O24...H12)	3.00 (O21...H12)	174.98	1.844
III	45.73(9) °	2.8 (O24...H14)	2.81 (O21...H14)	10.2	0.323

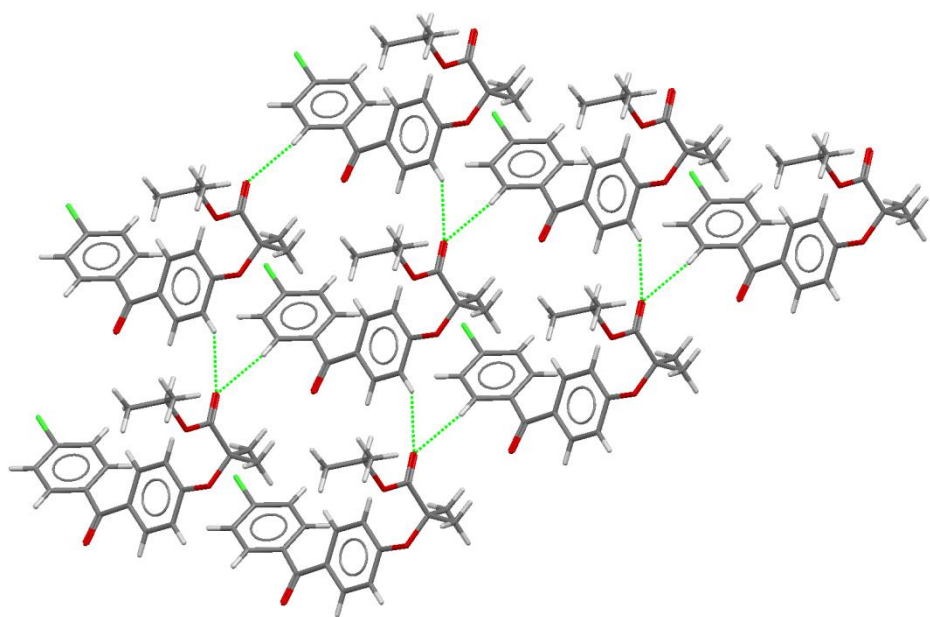
Table 8-2 further confirms the order of conformational similarity to the stable form I among the polymorphs to be form IIa > from III > from IIb in both the aromatic and aliphatic parts. The easiest way to see this is by comparing the values in the RMSD column. This value is calculated through the molecular similarity function in mercury, where the value obtained is based on the mean distance between all atoms in the different polymorphs. The values indicate that the molecular similarity of these forms is in the sequence mentioned earlier. Other parameters can also show the similarity between these molecules. For example, the angle between the planes of the two aromatic rings in form IIa deviates from the most stable form by less than 0.5 °, compared to a deviation of 3 ° in form III and 5 ° in form IIb. Similar trend was observed in the flexible part of the molecule as shown by the torsion angle between the central benzene ring and the propanyl-2-methylpropanoate group. The torsion angle between (C14-C13-O16-C17) was given in Table 8-2 as this is the angle at which the direction of the alkyl chain is defined; therefore it signals the similarity between the general structures of the different forms.

Form IIb adopts a conformation that is different from other known polymorphs, nonetheless, it was found that this molecule in this crystal has a similar conformation to the one observed in the fenofibric acid crystal structure.⁵ In the further discussion only the structures determined within this work (forms IIa and III) and the structure of the stable form I will be compared.

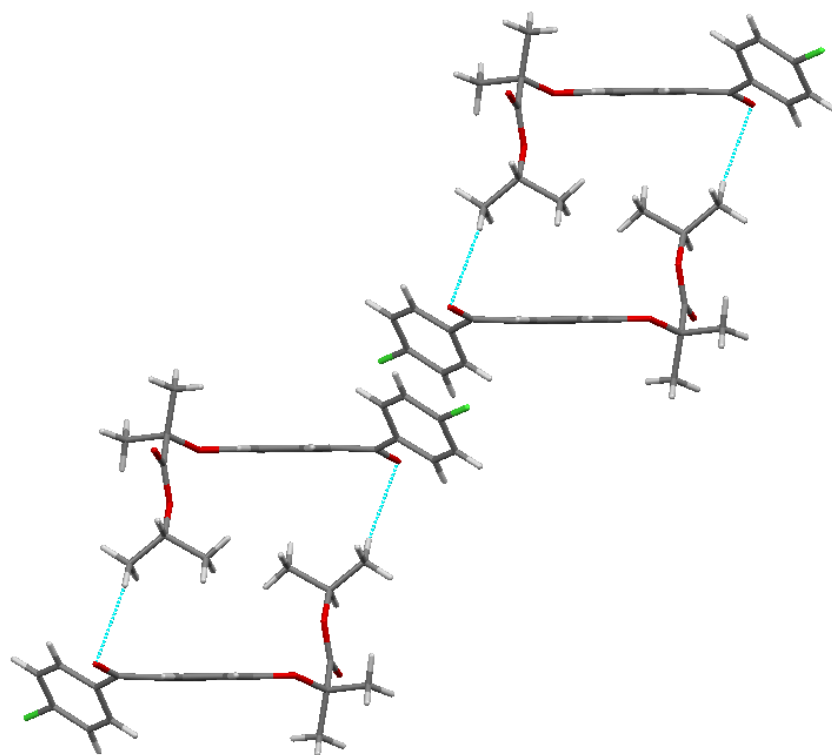
8.4.2 Main intermolecular interactions

Fenofibrate form I is the thermodynamically stable form meaning that it has the lowest Gibbs energy. For this reason, the interactions in the two newly found structures will be compared to it. The main groups that formed strong interactions in the investigated fenofibrate forms were the keto group between the two rings, the chlorbenzyl rings and the ester group in the aliphatic part. The interactions formed by these three groups will be discussed for each of the forms I, IIa and III.

In form I, the shortest intermolecular interaction was formed between the carbonyl oxygen (O24) of the ester group and a hydrogen atom (H12) of the phenyl ring [$d(\text{H}\cdots\text{O}) = (2.49 \text{ \AA})$, $\angle(\text{C-H}\cdots\text{O}) = 128^\circ$]. The same carbonyl atom forms another short interaction to the H5 atom of the chlorophenyl ring [$d(\text{H}\cdots\text{O}) = (2.61 \text{ \AA})$, $\angle(\text{C-H}\cdots\text{O}) = 172^\circ$], as illustrated in Figure 8-5(a). Both interactions are more than 0.1 Å shorter than the sum of the atomic van der Waals radii. These interactions connect fenofibrate molecules into a layer parallel to the (001) plane. The layers are connected to each other through further C–H \cdots O interactions, formed between isopropyl methyl groups (H25A) and ketone carbonyl oxygen (O9) atoms [$d(\text{H}\cdots\text{O}) = (2.66 \text{ \AA})$, $\angle(\text{C-H}\cdots\text{O}) = 156^\circ$] and offset π – π interactions between parallel chlorophenyl rings at an interplanar distance of 3.5116(6) Å, as seen in Figure 8-5(b). The interlayer C–H \cdots O interactions also facilitate the efficient packing of the isopropyl groups with a high surface area, which can be seen forming an “embrace” together.



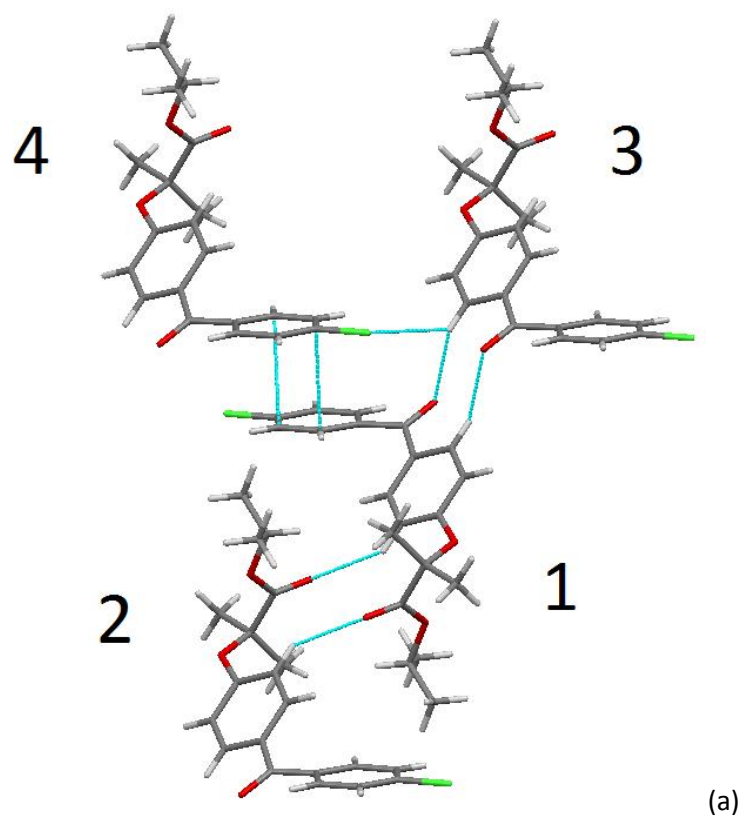
(a)



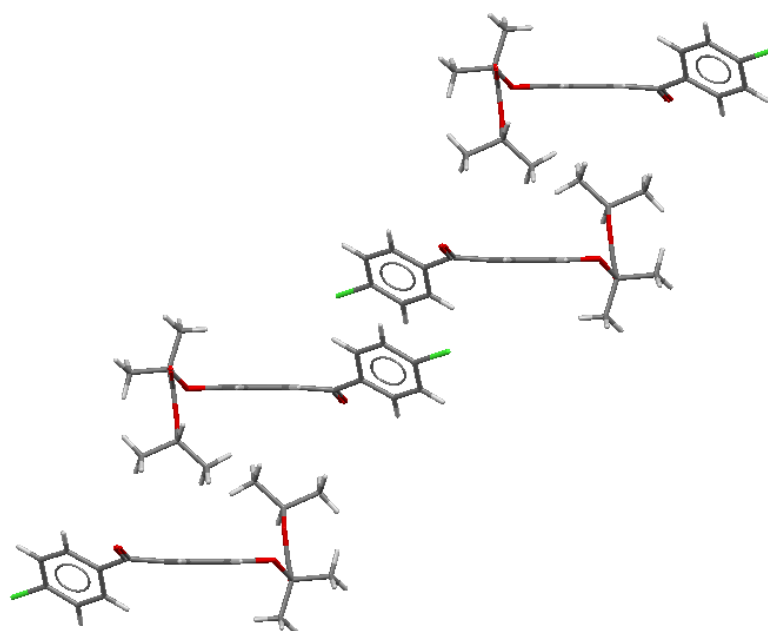
(b)

Figure 8-5. (a) The sheet that is formed parallel to the (001) plane *via* C-H...O interactions in Form I. (b) The offset π - π interactions between parallel chlorophenyl rings in addition to the "embrace" interaction in form I.

In form IIa, the carbonyl of the ester group in the aliphatic chain (O24) contributes to the formation of a dimer in the unit cell *via* C14–H14...O24 [$d(\text{H}\cdots\text{O}) = 2.662(11) \text{ \AA}$, $\angle(\text{C}–\text{H}\cdots\text{O}) = 129.6(8)^\circ$] and C15–H15...O24 [$d(\text{H}\cdots\text{O}) = 2.891(10) \text{ \AA}$, $\angle(\text{C}–\text{H}\cdots\text{O}) = 120.4(12)^\circ$] hydrogen bonds. These hydrogen bonds generate a ring motif (Figure 8-6(a), between molecules number 1 and 2). The large contact surface area between the two molecules suggests that van der Waals interactions play a significant role in stabilising the dimer. Two principal interactions connect adjacent dimers to form layers parallel with the (111) plane. These interactions are: C11–H11...O9 hydrogen bonds between the hydrogen donor from the benzene ring and the ketone carbonyl acceptor [$d(\text{H}\cdots\text{O}) = 2.570(16) \text{ \AA}$, $\angle(\text{C}–\text{H}\cdots\text{O}) = 135.2(8)^\circ$] and π – π interactions between chlorophenyl rings. The former forms a ring motif $R_2^2(14)$ between two molecules, as seen in Figure 8-6(a) (molecules 1 and 3). The offset π – π interaction involves the chlorobenzene fragments of two molecules and takes place at an interplanar distance of $3.3029(8) \text{ \AA}$ between the parallel rings, as seen in Figure 8-6(a) between molecules 1 and 4. The same stacking interaction was observed in form I with an interplanar distance of $3.5116(6) \text{ \AA}$. Moreover, the layers formed in this form also show the same “embrace” of the alkyl chain seen in form I, however at a larger inter molecular distance [Figure 8-6(b)].



(a)



(b)

Figure 8-6. (a) The main interactions in fenofibrate form IIa. (b) From IIa pattern that is common with form I.

In form III the molecules are linked into layers by two C–H···O bonds parallel to the crystallographic (100) plane. Specifically, one interaction takes place between a hydrogen of the isopropyl group (H23B) and the ketone carbonyl group of an adjacent molecule, O9 [$d(\text{H}\cdots\text{O}) = 2.56 \text{ \AA}$, $\angle\text{C–H}\cdots\text{O} = 160^\circ$]. The other interaction is between the same keto group (O9) and an isopropyl hydrogen (H18), from a different molecule [$d(\text{H}\cdots\text{O}) = 2.67(1) \text{ \AA}$, $\angle\text{C–H}\cdots\text{O} = 143.6^\circ$]. These interactions are illustrated in Figure 8-8(a). In this sheet arrangement there is a chain shared between forms I and III, but not present in form IIa [Figure 8-5(a) and Figure 8-8(a)]. Despite the similarity of the layer structures in forms I and III, their layers are not superimposable. The relative positions of the molecules perpendicular to the (100) plane of form III are different in the two forms.

The keto group in form III forms a C–H···O interaction with the hydrogen of benzene ring (H14) [Figure 8-8(b)]. This interaction forms a ring motif, resulting in a dimer. The C–H···O interaction takes place between the carbonyl of the keto-group (O9) and (H12) [$d(\text{H}\cdots\text{O}) = 2.92 \text{ \AA}$, $\angle\text{C–H}\cdots\text{O} = 161.8^\circ$]. It can be noticed that a repulsive H–H interaction is seen in this ring motif. Such repulsive interactions in the structure are thought to be a reason for the instability of this form. Similar interaction between the keto group and a central benzene hydrogen was observed in form IIa [Figure 8-6(a), molecules 1 and 3], but not in form I.

Form III shares two other interactions with both forms I and IIa. Firstly, the offset π – π interaction of the chlorobenzene rings. In form III, this interaction links pairs of dimers at a distance of $3.5719(8) \text{ \AA}$. The second shared interaction is the “embrace” type of interaction, this seem to be the most essential interaction in all forms, where it maximizes the contact surface of molecules in the crystal. The ether atom O(21) has a short contact to H23 ($2.78(2) \text{ \AA}$) and between carbonyl O24 and H atoms of the chlorobenzyl ring (H5 and H6; 2.84 \AA and 2.88 \AA). The offset π – π interaction, the “embrace” interaction and the ether C–H···O interaction can be seen in Figure 8-8(c) and (d).

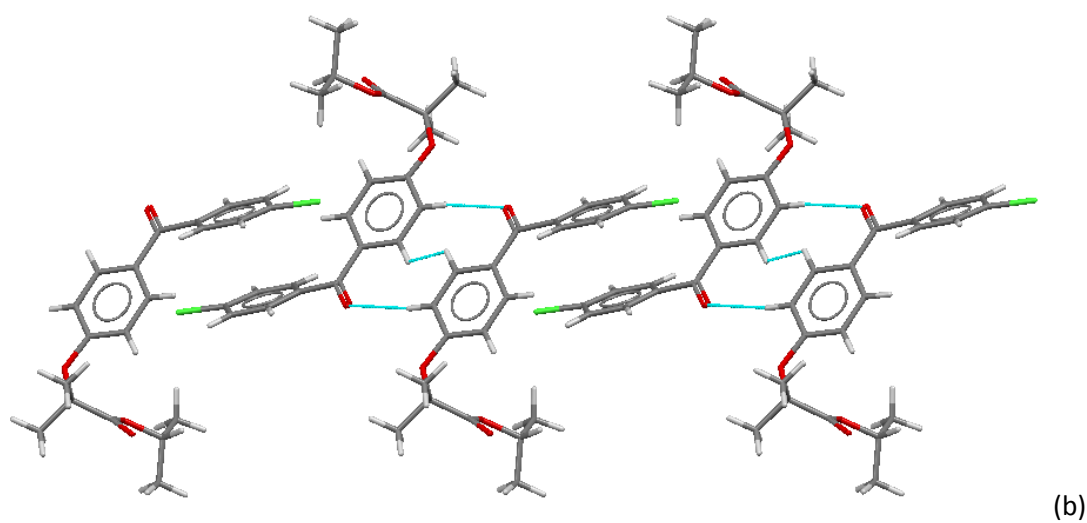
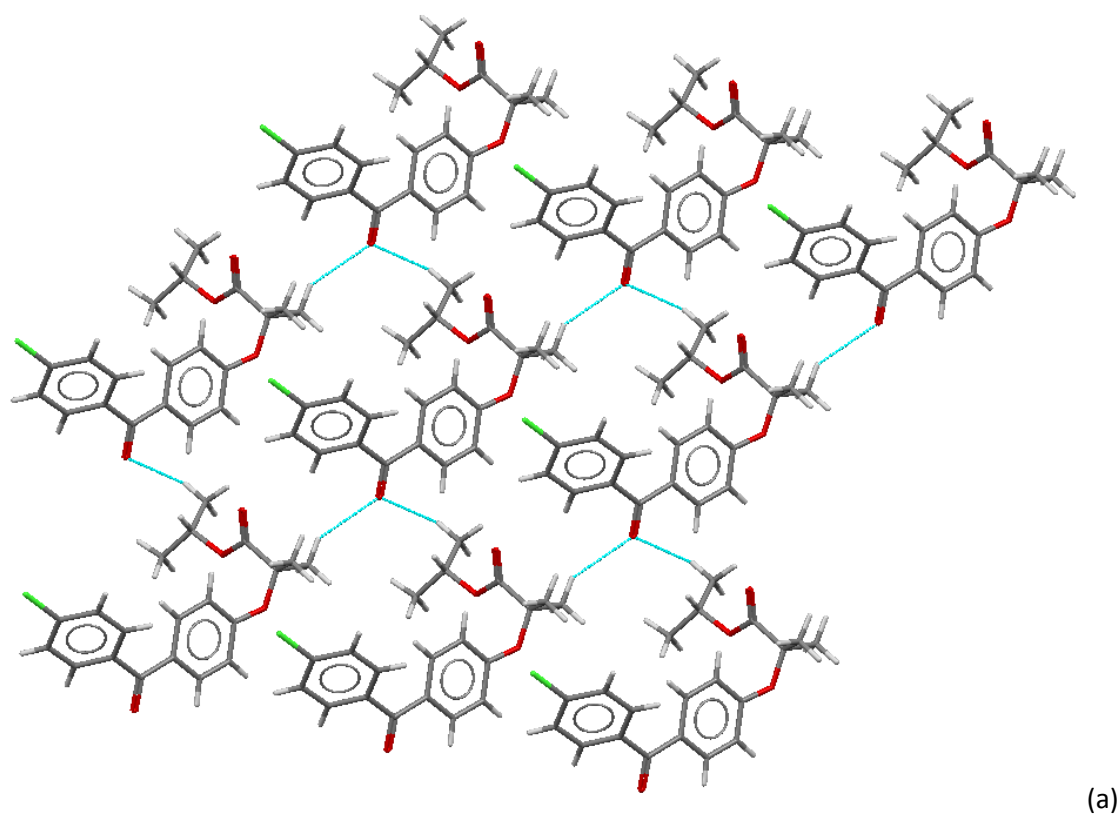
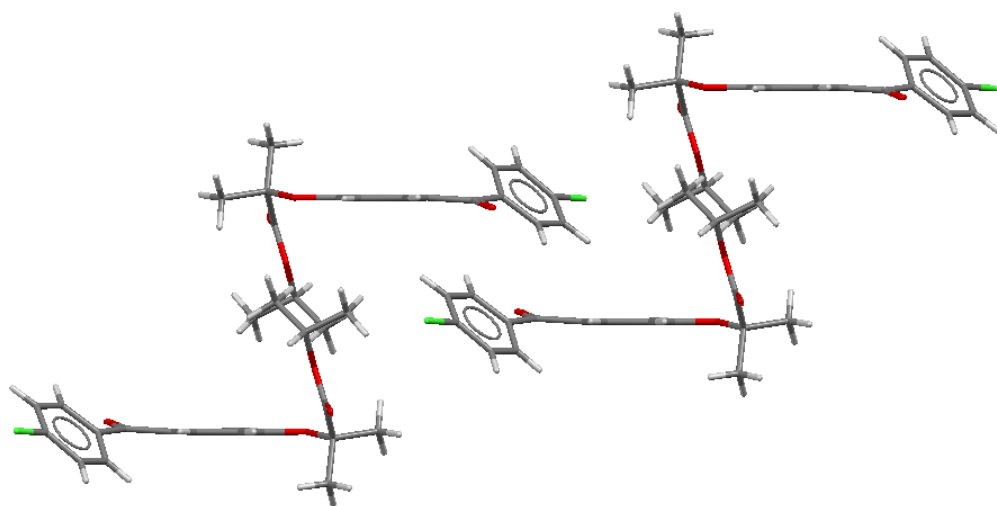
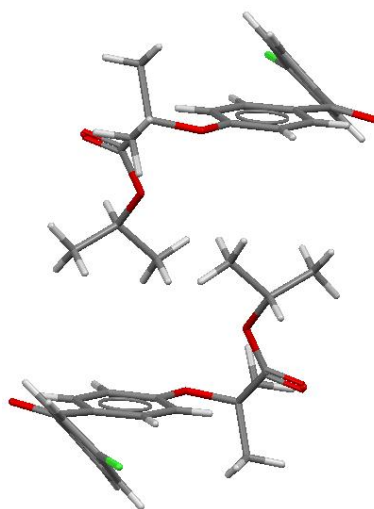


Figure 8-7. (a) The sheet that is formed parallel to the (100) plane *via* C-H...O interactions in form III. (b) The keto group in C-H...O interaction with the hydrogen of benzene ring in form III.



(c)



(d)

Figure 8-8. Continued. (c) Illustrates the offset π - π interaction, the “embrace” interaction and the ether C-H \cdots O interaction in form III. (d) The two molecules on the right side in part (c) from a different angle, showing the embrace type of interaction in form III, not short H-H interaction is noticed between the two molecules in part (d).

8.4.3 Comparison of crystal structures of fenofibrate form I, form IIa and form III

A summary of the relevant interactions in the three forms of fenofibrate, I, IIa and III, is given in Table 8-3.

Table 8-3. Relevant short H-acceptor interactions in the crystal structures of fenofibrate forms I, IIa and III. The H atoms have the same numbers as C atoms that they attached to. (See Figure 8-1 for atom numbering). The symbols A, H and D in the table are short of Acceptor, Hydrogen and Donor atoms

A	H	D	A...H distance	D...A distance	D-H	Angle	vdW difference	Symmetry
<i>Form I</i>								
O24	H12	C12	2.49	3.156(2)	0.93	128.4	-0.23	$x,1+y,z$
O24	H5	C5	2.61	3.534(2)	0.93	171.8	-0.11	$1+x,1+y,z$
O24	H11	H11	2.78	3.296(2)	0.93	116.4	0.06	$x,1+y,z$
O9	H18A	C18	2.62	3.523(2)	0.96	156.3	-0.1	$-1+x,-1+y,z$
O9	H19A	C19	2.87	3.702(2)	0.960	145.4	0.15	$-1+x,-1+y,z$
O9	H2	C2	2.91	3.171(2)	0.930	98.0	0.19	$2-x,-y,1-z$
O9	H23A	C23	2.67	3.401(3)	0.960	133.8	-0.06	$3-x,1-y,2-z$
Cl7	H14	C14	3.10	3.706(2)	0.93	124.4	0.15	$2-x,1-y,1-z$
<i>Form IIa</i>								
Cl7	H11	C11	2.895(13)	3.541(2)	0.96(2)	125.8(9)	-0.06	$1+x,-1+y,z$
Cl7	H6	C6	2.924(18)	3.720(2)	0.96(2)	141(1)	-0.03	$3-x,1-y,-z$
Cl7	H25A	C25	3.062(12)	3.967(3)	0.98(1)	155(1)	0.11	$2-x,1-y,1-z$
O9	H11	C11	2.570(16)	3.318(3)	0.96(2)	135.2(9)	-0.15	$1-x,2-y,-z$
O9	H5	C5	2.874(12)	3.571(3)	0.95(3)	131.2(9)	0.15	$2-x,2-y,-z$
O16	H18A	C18	2.773(15)	3.734(3)	0.97(2)	170.6(9)	0.05	$-x,2-y,1-z$
O24	H14	C14	2.662(11)	3.374(2)	0.98(2)	129.6(9)	-0.06	$1-x,1-y,1-z$
O24	H18B	C18	2.783(10)	3.533(3)	0.97(1)	134.5(8)	0.06	$-x,1-y,1-z$
O24	H19A	C19	2.913(15)	3.697(3)	0.99(1)	136.9(9)	0.19	$-x,1-y,1-z$
O24	H15	C15	2.891(10)	3.485(3)	0.97(2)	120(1)	0.17	$1-x,1-y,1-z$

Table 8-3. Continued

Form III								
Cl7	H23A	C23	3.114(15)	3.915(3)	1.01(2)	141(1)	0.16	1-x,1-y,2-z
Cl7	H18A	C18	2.960(11)	3.781(2)	0.98(1)	141.9(8)	0.01	1+x,1+y,1+z
Cl7	H25A	C25	2.828(14)	3.586(3)	0.98(1)	135.1(8)	-0.12	1+x,1+y,1+z
O9	H18B	C18	2.671(12)	3.509(2)	0.98(1)	143.5(8)	-0.05	x,1+y,1+z
O9	H25B	C25	2.565(12)	3.498(4)	0.98(1)	159.9(9)	-0.15	x,y,1+z
O9	H12	C12	2.92	3.810(2)	0.93	161.8	0.2	-x,1-y,2-z
O24	H22	C22	2.91	3.772(2)	0.98	146.9	0.19	1-x,1-y,1-z
O24	H6	C6	2.88	3.468(2)	0.93	122.2	0.16	x,-1+y,-1+z
O24	H5	C5	2.84	3.449(2)	0.93	124.4	0.12	x,-1+y,-1+z
O21	H23C	C23	2.775(16)	3.660(3)	1.01(2)	147(1)	0.06	-x,1-y,1-z

Interestingly, no interactions were observed at a distance less than the Van der Waal radii of interacting atoms seem to be shared between the three forms. However, some similarities can be identified between forms I and III. For example, the O9 of the ketone group interacts with the methyl group (O9 ...H18B) and with isopropyl group (O9 ...H25B). The interaction between O24 of the ester carbonyl and H23 of the isopropyl group is also common for both forms I and III. Despite these differences in the interactions among the three fenofibrate forms, similar packing trends can be identified. Each molecule of the three forms adopts a capital “L” letter shape. Two of these “L”s pack inverted to each other to form a near-rectangular shape “L7”. Rectangular dimers then efficiently fill the space. An illustration of the packing of the three forms is shown in Figure 8-9.

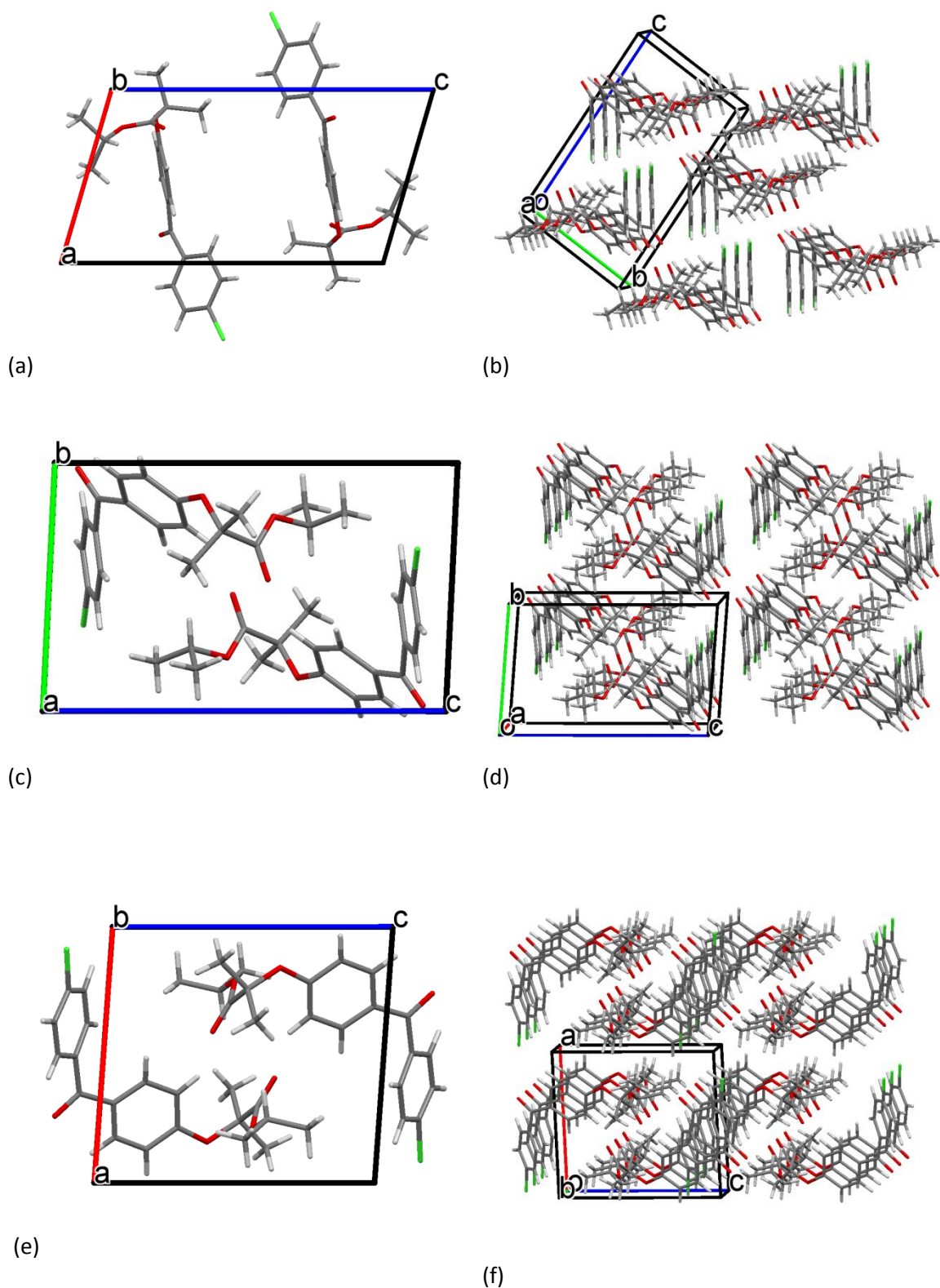


Figure 8-9. The unit cell and packing of fenofibrate form I (a) and (b), form IIa (c) and (d) and form III (e) and (f).

In order to know how similar these packing patterns are to each other, the “crystal packing similarity” function in Mercury was used to compare forms IIa and III to the most stable form I. The packing of 15 molecules in each form were compared. Form IIa had 6 out of 15 molecules

in common with form I compared to 1 molecule out of 15 in common between form III and form I. The observation that the packing is more similar for forms I and IIa could be an explanation to the experimentally observed sequence of polymorph transitions: III->IIa->I when fenofibrate is left to age at room temperature. In order to assume the conformation and arrangement of form I, molecules in form III first need to convert to the other metastable polymorph IIa which then can undergo transition to form I.

8.5 References

1. Amstad E, Spaepen F, Weitz DA. Crystallization of undercooled liquid fenofibrate. *Physical Chemistry Chemical Physics*. 2015;17(44):30158-61.
2. Yang Z. Development of Methods to Predict and Enhance the Physical Stability of Hot Melt Extruded Solid Dispersions: University of East Anglia; 2013.
3. Di Martino P, Palmieri GF, Martelli S. Evidence of a metastable form of fenofibrate. *Die Pharmazie*. 2000;55(8):625-6.
4. Heinz A, Gordon KC, McGoverin CM, Rades T, Strachan CJ. Understanding the solid-state forms of fenofibrate – A spectroscopic and computational study. *European Journal of Pharmaceutics and Biopharmaceutics*. 2009;71(1):100-8.
5. Balendiran GK, Rath N, Kotheimer A, Miller C, Zeller M, Rath NP. Biomolecular Chemistry of Isopropyl Fibrates. *Journal of Pharmaceutical Sciences*. 2012;101(4):1555-69.
6. Górniak A, Wojakowska A, Karolewicz B, Pluta J. Phase diagram and dissolution studies of the fenofibrate–acetylsalicylic acid system. *Journal of Thermal Analysis and Calorimetry*. 2011;104(3):1195-200.
7. Watterson S, Hudson S, Svärd M, Rasmuson ÅC. Thermodynamics of fenofibrate and solubility in pure organic solvents. *Fluid Phase Equilibria*. 2014;367:143-50.
8. Tipduangta P, Takieddin K, Fábíán L, Belton P, Qi S. A New Low Melting-Point Polymorph of Fenofibrate Prepared via Talc Induced Heterogeneous Nucleation. *Crystal Growth & Design*. 2015;15(10):5011-20.
9. Henry RF, Zhang GZ, Gao Y, Buckner IS. Fenofibrate. *Acta Crystallographica Section E*. 2003;59(5):o699-o700.

Chapter 9: Summary and conclusions

9.1 Summary

Pharmaceutical companies are always trying to find new solid forms of APIs for different reasons. For example new solid forms have the potential to give new forms that give superior physicochemical properties to the known forms of a drug such as stability and dissolution (as in the case of amoxicillin). Additionally, a newly discovered form is not covered by the patent of the parent/innovator company, allowing for large profits of the company discovering the new form. One of the strategies to find new solid forms is the use of multicomponent solids. This method has gained popularity over the past few years. The forms of cocrystal and solvate have received special attention because they are less well explored. The number of marketed multicomponent drugs is certainly increasing, showing the larger interest in this area. These forms can also have negative effects on pharmaceutical companies, where the unexpected formation of a solvate could lead to having remnants of the solvents used during processing of the API, causing a danger to the patient as well as a financial loss to the company.

Until today, the formation of multicomponent solids is not readily predictable, so high-throughput screening methods are used in order to cover the possible range of solid forms (single- and multi-component) that an API can form. Crystal structure prediction is also being used alongside experimental work in order to show the possible forms using lattice energy calculations.

In this work, we focused on the solvate forms of organic materials. We tried to relate the structural features of organic compounds to their ability to form a solvate using a knowledge-based approach. Identifying these features resulted in better understanding of the crystallization of these forms and provides a quick, easy-to-use tool for scientists, helping them in the choice of solvents in early stage development.

As a proof of concept, four of the most commonly used recrystallization solvents for organic materials were studied. These solvents were ethanol, methanol, dichloromethane, chloroform. Additionally, water, being the preferred choice in any experiment, was studied as well. The Cambridge Structural Database, is the only database that contains the relevant data.

For each of the five solvents, two groups of entries were extracted from the CSD; these are a solvate-forming and a non-solvate-forming group. For each extracted molecule, thousands of molecular descriptors were calculated *via* the Dragon software. The output of Dragon was 10 tables; a solvate and a non-solvate forming table per solvent. Each of these tables had a few hundreds to a few thousands molecules, which were represented by rows and had thousands of descriptors, represented by columns of these tables. All of these descriptors were numerical values, which makes them suitable for the application of statistical procedures.

Statistical significance testing was applied to each solvent's dataset individually, comparing the solvate and non-solvate forming groups. The majority of the descriptors showed significant difference between solvate and non-solvate molecules. This implies that most of the structural features represented by molecular descriptors have shown the potential to be useful in differentiating the solvate and the non-solvate molecules.

Due to the large number of these useful descriptors, machine learning algorithms were used. Initially, an unsupervised machine learning method, PCA, was applied to the data in order to reduce the dimensionality. The PCA gave a good separation of the data in space. Nevertheless, PCA could not point towards a small number of descriptors that could be used to achieve the good separation that was seen by PCA. After that, supervised machine learning was used to learn from the available data. A linear and non-linear method were compared, these are logistic regression and support vector machine, respectively. The performance of these two methods was comparable, with a higher consistency in logistic regression performance,

implying that support vector machines could be overfitting to the training data in some cases. Therefore logistic regression was used for the rest of the analysis.

Logistic regression was first used to fit models using the first, first two and first three principal components (PCs). It was then used in a systematic approach to descriptor selection, that is, models with all possible combinations of one and two descriptors were fitted. When the results were compared, the models with one and two variables had comparable results with the PCA logistic regression models. This could reflect the high correlation between descriptors, where the dimensionality reduction was not of great significance. The variables with two descriptors slightly outperformed PCA in the models of all five solvents. Additionally, they showed exactly what chemical features are the main contributors for solvate formation, in terms of the descriptors available. A third descriptor was added to the models using forward selection but it couldn't improve the performance in any of the five solvents. This again shows the high correlation between variables, where among thousands of potentially useful descriptors, the predictive ability stopped improving when the third descriptor was added. For this reason, the two-variable models were considered to be optimal. The two variables in these models were related to the size and branching of molecules in addition to the availability of heteroatoms in a molecule. When these models were evaluated, a small bias was noticed, with a preferential prediction towards the non-solvate group in all solvents. This was corrected by adjusting the intercept of each model, until the bias was not seen any further. The next step that was taken was to provide a simpler description than the one given by the two-variable models. This could be useful for researches who want to make predictions without using computers. The simpler models were obtained based on the correlation of simple descriptors to the ones in the two-variable models. As expected, the resulting simple models performed very closely to the two-variable models.

Analysis of correctly and incorrectly predicted structures revealed that the importance of size and branching of the molecule exceeds any other factor, at least among the studied molecular descriptors. Factors that seem to be involved in solvate formation but the models did not take them into account (e.g. C–H...O hydrogen bonding, steric hindrance and halogen bonding) were also identified. These factors are thought to be the reason for misprediction in many cases, but were too few in number to be found by the machine learning algorithms. This assumption was supported by examples from each data set showing mispredictions caused by ignoring at least one of these factors.

The applicability of the models to pharmaceuticals was tested *via* experimental validation. Ten drug compounds that vary in their properties, especially in the factors that the models take into account, were used in this validation. Each of the candidates was tested for its ability to form a solvate with each of the five solvents, and the results of the experimental screening were compared to the predictions. The success rates for ethanol and methanol were the highest (9/10) followed by dichloromethane and chloroform (8/10) then followed by the water model (5/10). Multiple reasons could have caused these results. For example the number of points in the training data was the smallest in the water data set; therefore it is expected for the water model to have the lowest predictability. Additionally, a test set of 10 drug candidates is small. Consequently, the test set was probably not representative of the population of known crystal structures. Therefore the validation was more of a demonstration of how the models work with pharmaceutically active materials and how realistic the results are giving probabilities ranging between 0 and 1 depending on the structure given.

During the screening of the solvate formation, some of these drug candidates formed solvate forms that were not previously reported. In this case, the PXRD of the sample was collected. Additionally, a new solvate form of griseofulvin dichloromethane solvate was reported *via* SCXRD experiment. A dichloromethane solvate of griseofulvin has previously been reported.

Nevertheless, the crystallographic parameters were different from the reported form, showing an example of solvate polymorphism.

Chapter 8 was my contribution to a study conducted by P. Tipduangta, a colleague in the same suit who is studying the heterogeneous crystallization of fenofibrate, an anti-hyperlipdemic agent. The study focused on the effect of surface annealing and temperature in preferential crystallization of different polymorphic forms, where it was possible to obtain polymorph IIa of fenofibrate purely by manipulating these factors. It also focused on the role of a commonly used additive, talc, in acquiring a new polymorph of fenofibrate that was not previously reported, which was denoted as polymorph III. The crystal structures of forms IIa and III had not been previously reported. The main scope of this chapter was the determination and analysis of the new structures. The structural similarities and differences between the newly obtained and the known polymorphic forms were highlighted and discussed in terms of possible transition pathways.

9.2 Conclusions and future outlook

9.2.1 Conclusions

The work produced in this thesis has identified the main structural features that are responsible for solvate formation with five solvents in organic compounds using a knowledge-based approach. These features were used to fit one predictive model per solvent. The models can be used to predict the probability of solvate formation of any organic compound with each of the five solvents. The factors that were found to influence solvate formation were mainly related to size and branching of a molecule, along with its hydrogen bonding ability. The success rate obtained ranged between 74-80 % in all 5 solvents. This shows the potential for cheminformatic approaches to become one of the fundamental methods in many steps of drug discovery and development, providing a simple, quick guide for scientists. Implementation of cheminformatics in gaining knowledge and predicting future results can save a lot of time and money in the drug development cycle.

The key points learned from the project will can be summarized in the following points:

(1) Careful collection and pre-processing of data are important steps, where consideration of one phenomenon, such as polymorphism, before analysing the data has resulted in dropping more than 30 % of the parent data due to redundancy. Redundancy isn't a problem with data only, but also applies to the description of data, where most of the ~5000 descriptors used turned out to have high correlation to each other, resulting in redundant information.

(2) A linear, simple algorithm was enough for preliminary classification in this problem. A more complex, non-linear method has resulted in overfitting to the training data.

(3) Principal components have reduced the dimensionality of the data, but failed to add value, possibly due to the high correlation between descriptors.

(4) MSE, AIC and AUC gave very similar results when they were used for model selection in this problem, indicating their interchangeability.

(5) Complex descriptors are not an essential requirement for models with a good predictability, at least for a preliminary description of the data. In this problem complex descriptors could be replaced by simple, correlated descriptors, resulting in comparable predictive ability.

(6) Statistical methods are subject to overfitting and underfitting, depending on the sensitivity on the method applied. Therefore the consideration of the chemical structure or the opinion of an expert is required for a valid prediction.

(7) The size of a validation set is important to be large enough to represent the data, 10 drug candidates per solvent did not provide enough data to confidently judge the experimental validity of the predictive models.

(8) Crystallization conditions, such as exposed surface area and temperature can result in different polymorphic forms. Although the last piece of work in this thesis did not explicitly concern solvates, crystallization of different forms have the potential to be applied in the field of multicomponent solids.

9.2.2 Future outlook

The work presented in this thesis has proven the usefulness of cheminformatics approaches in predicting the physicochemical behaviour of organic compounds. While the prediction isn't perfect, improving such a predictive method is possible. One of the main requirements for obtaining a better predictive ability is collection of more data. Machine learning is based upon using the training data to fit predictive models. The bigger the training data and the more diverse it is, the more likely it is to give a realistic model. Another way to improve these models would be finding a better description of compounds. This includes a more detailed description of hydrogen, halogen and π interactions. It is also possible to add new descriptions, encoding new information that is not yet encoded in descriptors such as the accessibility of atoms, based on the molecular graph. This could be useful especially when atoms with the potential to form strong intermolecular bonds exist.

The successful application of knowledge-based systems in a complex problem, such as solvate formation indicates the possibility to extend their use in other phenomena in the solid state, such as cocrystal formation, drug-polymer compatibility and stability of solid dispersions. It is important to note here that in order to obtain meaningful results from knowledge-based methods, chemical knowledge of a problem is required. Such methods serve as a guide for researchers, where it could point towards the factors that need to be considered in their research.